Technical Report 717

Project A:
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

Development and Field Test of Task-Based MOS-Specific Criterion Measures

Charlotte H. Campbell and Roy C. Campbell
Human Resources Research Organization

Michael G. Rumsey
Army Research Institute

Dorothy C. Edwards
American Institutes for Research

Selection and Classification Area
Manpower and Personnel Research Laboratory

U.S. Army

Research Institute for the Behavioral and Social Sciences

July 1986

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the
Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Jane M. Arabian
Paul G. Rossmeissl

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A182 695

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 717
6a. NAME OF PERFORMING ORGANIZATION Human Resources Research Organization 6b. OFFICE SYMBOL (If applicable) HumRRO	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences (ARI)
6c. ADDRESS (City, State, and ZIP Code) 1100 South Washington Street Alexandria, VA 22314-4499	7b. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA 903-82-C-0531
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. 20263731 A792 / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Development and Field Test of Task-Based MOS-Specific Criterion Measures

12. PERSONAL AUTHOR(S) Campbell, Charlotte C. (HumRRO), Campbell, Roy C. (HumRRO), Rumsey, Michael G. (ARI), & Edwards, Dorothy C. (American Institutes for Research)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
	FROM Oct 83 TO Oct 85	July 1986	80

16. SUPPLEMENTARY NOTATION

The Army Research Institute technical point of contact is Dr. Lawrence M. Hanser. His telephone number is (202) 274-8275.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Classification, Criterion Measures, Hands-on Tests, Job Experience, Knowledge Tests, MOS-Specific Tests, Performance Ratings, Project A Field Test, Selection, Soldier Effectiveness

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The research described in this report was performed under Project A, the U.S. Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This report documents the development and field tryout of task-based MOS-specific knowledge tests, hands-on tests, task performance ratings, and job experience questionnaires for nine Military Occupational Specialties (MOS). Job performance domains were derived from Army Occupational Survey Programs (AOSP) results, Soldier's Manuals, and proponent agency input. Subject-matter expert judgments of task criticality, difficulty, and similarity were used to select tasks for test development. All tests were pilot tested on Skill Level 1 soldiers and noncommissioned officers. Field tests were conducted among 114-178 soldiers per MOS. Results were used to revise the instruments and to provide evidence of reliability and validity. Proponent agencies provided technical reviews before the field tests and after the instruments were revised.

(continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Lawrence M. Hanser	22b. TELEPHONE (Include Area Code) (202) 274-8275	22c. OFFICE SYMBOL PERI-RS

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

ARI Technical Report 717

19. (Continued)

Hands-on and knowledge tests exhibited reasonable performance variability, as did the rating scales. Correlations between the two test methods were high, but do not suggest that either should substitute for the other. Rating scales correlated more highly among themselves than with the tests, as would be expected from their surface dissimilarity and affective focus. Job experience emerged as a potentially important factor in explaining performance variability.

The instruments were finalized for the upcoming Concurrent Validation, where they will serve as criterion measures for a new predictor battery designed to supplement the Armed Services Vocational Aptitude Battery (ASVAB).

The appendixes that present the documentation for this research are contained in ten separate volumes, Appendixes to ARI Technical Report 717, Development and Field Test of Task-Based MOS-Specific Criterion Measures.

Volume 1: Appendixes A-E.

Volume 2: Appendixes F and G (Part 1) (limited distribution)

Volume 3: Appendix G (Part 2) (limited distribution)

Volume 4: Appendix H (Part 1) (limited distribution)

Volume 5: Appendixes H (Part 2), I, and J (limited distribution)

Volume 6: Appendixes K, L, M, N, O, P, Q, R, and S

Volume 7: Appendixes T, U, and V (Part 1) (limited distribution)

Volume 8: Appendix V (Part 2) (limited distribution)

Volume 9: Appendix V (Part 3) (limited distribution)

Volume 10: Appendix V (Part 4) (limited distribution)

See Table of Contents for more detailed listing of the contents of each of the appendixes.

Technical Report 717

Project A:
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

Development and Field Test of Task-Based MOS-Specific Criterion Measures

Charlotte H. Campbell and Roy C. Campbell
Human Resources Research Organization

Michael G. Rumsey
Army Research Institute

Dorothy C. Edwards
American Institutes for Research

Selection and Classification Technical Area
Lawrence M. Hanser, Acting Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

July 1986

Army Project Number
2Q263731A792

Manpower and Personnel

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

This document describes the development and field testing of task-based MOS-specific criterion measures for evaluating the performance of Army enlisted personnel. The research was part of the Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance. The portion of the effort described herein is devoted to the development and validation of Army Selection and Classification Measures, and referred to as "Project A." Another part of the effort is the development of a prototype Computerized Personnel Allocation System, referred to as "Project B." Together, these Army Research Institute efforts, with their in-house and contract components, comprise a major program to develop a state-of-the-art, empirically validated system for personnel selection, classification, and allocation.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

As should be obvious from the length of this report, many people were involved in the development and field-testing of the MOS-specific criterion measures. The major contributors, in addition to the authors, included (in alphabetical order):

Instrument Development

Kathy Cordray-Austin (AIR)
Jack Doyle (HumRRO)
Daniel Felker (AIR)
Pat Ford (HumRRO)
Gene Hoffman (HumRRO)
Paul Radtke (AIR)
Aurelia Scott (AIR)

Instrument Preparation

Candace Hoffman (HumRRO)
Mary Lodge (AIR)
Deborah Marcum (HumRRO)

Field Test Support

James H. Harris (HumRRO)
LTC (Ret.) Donald Rogan (ARI)

Data Analysis

Laurie Wise (AIR)
Winnie Young (AIR)

Report Guidance

John P. Campbell (HumRRO)
Lola Zook (HumRRO)

DEVELOPMENT AND FIELD TEST OF TASK-BASED MOS-SPECIFIC CRITERION MEASURES

EXECUTIVE SUMMARY

Requirement:

The purpose of this report is to describe the activities involved in the development and field tryout of job-specific, task-based criterion measures for nine Army MOS. The measures include hands-on tests, knowledge tests, performance ratings, and a job-experience questionnaire.

Procedure:

For each MOS, the job-performance domain was derived through examination of the Army Occupational Survey Program (AOSP) results, the Soldier's Manual of Common Tasks, MOS-specific Soldier's Manuals, and MOS proponent agency input. Subject Matter Experts (SME) provided judgments of task criticality, difficulty, and similarity, and separate panels of SME in each MOS used those judgments to select 30 MOS tasks for development of performance measures.

For each MOS, knowledge (paper-and-pencil) tests and the questionnaire were written to cover all 30 tasks, while hands-on tests and rating scales were constructed for 15 of the tasks. Hands-on and knowledge tests were pilot-tested on small groups of soldiers and SME. The subsequent field tests of all instruments were conducted at eight locations, involving 114-178 soldiers per MOS. Results were used to revise the instruments as well as to provide evidence of their reliability and validity. MOS Proponent Agencies provided a technical review of the instruments before the field tests, and again after the instruments were revised on the basis of field test results. The report and appendixes document all steps in the development process and provide complete analysis summaries.

Findings:

In general, the hands-on and knowledge tests exhibited reasonable levels of performance variability. Correlations between the two methods indicate a high degree of relationship, but were not so high as to suggest that either may be a substitute for the other. While estimates of internal consistency were not high relative to other investigations, that is not a cause for alarm, given that the development strategy emphasized comprehensive coverage of heterogeneous jobs, rather than content homogeneity. Rating scales were more highly intercorrelated among themselves than with other measures, with high reliabilities. Because they are addressing the affective component of performance, rather than technical skill or knowledge, these findings argue for their retention in subsequent data collection. The Job History Questionnaire results suggest that experience, thus measured may be an important factor in explaining performance variability.

Utilization of Findings:

The job-specific task-based measures--hands-on tests, knowledge tests, task performance rating scales, and job experience questionnaire--are to be used as criterion measures against which a new predictor battery will be validated. They are to be administered, along with training achievement measures, MOS-nonspecific (Army-wide) performance rating scales, and the predictors, to large samples (about 650) of incumbents in each MOS in a concurrent validation design during FY 85.

DEVELOPMENT AND FIELD TEST OF TASK-BASED MOS-SPECIFIC CRITERION MEASURES

CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDIXES*

*The appendixes for this report are issued in separate reports. Those so designated have limited distribution.

CONTENTS (Continued)

CONTENTS (Continued)

CONTENTS (Continued)

DEVELOPMENT AND FIELD TEST OF TASK-BASED MOS-SPECIFIC CRITERION MEASURES

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program which the U.S. Army has undertaken to develop an improved personnel selection and classification system for enlisted personnel. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through use of new and improved selection/classification tests which will validly predict carefully developed measures of job performance. The project addresses the 675,000-person enlisted personnel system of the Army, encompassing several hundred different military occupations.

This research program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research effort that would be needed to develop the desired system. In 1982 a consortium led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI) was selected by ARI to undertake the 9-year project. The total project utilizes the services of 40 to 50 ARI and consortium researchers working collegially in a variety of specialties, such as industrial and organizational psychology, operations research, management science, and computer science.

The specific objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria. The latter are to include both Army-wide job performance measures based on newly developed rating scales, and direct hands-on measures of MOS-specific task performance.

- Develop and validate new selection and classification measures.

- Validate intermediate criteria (e.g., performance in training) as predictors of later criteria (e.g., job performance ratings), so that better informed reassignment and promotion decisions can be made throughout a soldier's career.

- Determine the relative utility to the Army of different performance levels across MOS.

- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The research design for the project incorporates three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first

iteration, file data from Army accessions in fiscal years (FY) 1981 and 1982 were evaluated to explore the relationships between the scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and their subsequent performance in training and their scores on the first-tour Skills and Qualification Tests (SQT).

In the second iteration, a concurrent validation design will be executed with FY83/84 accessions. As part of the preparation for the Concurrent Validation, a "preliminary battery" of perceptual, spatial, temperament/ personality, interest, and biodata predictor measures was assembled and used to test several thousand soldiers as they entered in four Military Occupational Specialties (MOS). The data from this "preliminary battery sample" along with information from a large-scale literature review and a set of structured, expert judgments were then used to identify "best bet" measures. These "best bet" measures were developed, pilot tested, and refined. The refined test battery was then field tested to assess reliabilities, "fakability," practice effects, and so forth. The resulting predictor battery, now called the "Trial Battery," which includes computer-administered perceptual and psychomotor measures, will be administered together with a comprehensive set of job performance indices based on job knowledge tests, hands-on job samples, and performance rating measures in the Concurrent Validation.

In the third iteration (the Longitudinal Validation), all of the measures, refined on the basis of experience in field testing and the Concurrent Validation, will be administered in a true predictive validity design. About 50,000 soldiers across 20 MOS will be included in the FY86-87 "Experimental Predictor Battery" administration and subsequent first-tour measurement. About 3500 of these soldiers are estimated for availability for second-tour performance measurement in FY91.

For both the concurrent and longitudinal validations, the sample of MOS was specially selected as a representative sample of the Army's 250+ sntry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These MOS account for about 45 percent of Army accessions. Sample sizes are sufficient so that race and sex fairness can be empirically evaluated in most MOS.

Activities and progress during the first two years of the project were reported for FY83 in ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37, and for FY84 in ARI Research Report 1393 and its related reports, ARI Technical Report 660 and ARI Research Note 85-14. Other publications on specific activities during those years are listed in those annual reports. The annual report on project-wide activities during FY85 is under preparation.

For administrative purposes, Project A is divided into five research tasks:

Task 1 -- Validity Analyses and Data Base Management
Task 2 -- Developing Predictors of Job Performance

Task 3 -- Developing Measures of School/Training Success
Task 4 -- Developing Measures of Army-Wide Performance
Task 5 -- Developing MOS-Specific Performance Measures

The development and revision of the wide variety of predictor and criterion measures reached the stage of extensive field testing during FY84 and the first half of FY85. These field tests resulted in the formulation of the test batteries that will be used in the comprehensive Concurrent Validation program which is being initiated in FY85.

The present report is one of five reports prepared under Tasks 2-5 to report the development of the measures and the results of the field tests, and to describe the measures to be used in Concurrent Validation. The five reports are:

Task 2 -- Development and Field Test of the Trial Battery for Project A, Norman G. Peterson, Editor, ARI Technical Report (in preparation).

Task 3 -- Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS, by Robert H. Davis, et al., ARI Technical Report (in preparation).

Task 4 -- Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program, by Elaine D. Pulakos and Walter C. Borman, Editors, ARI Technical Report 716.

Task 5 -- Development and Field Test of Task-Based MOS-Specific Criterion Measures, Charlotte H. Campbell, et al., ARI Technical Report 717.

-- Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS, Jody L. Toquam, et al., ARI Technical Report (in preparation).

Chapter 1

INTRODUCTION

CONTEXT

The overall purpose of Project A is to develop an expanded and comprehensive selection/classification test battery and validate that battery against a full array of existing and project-developed criteria. The work reported in this document was directed toward developing the criteria for validating the test battery and deals with task-based job-specific measures.

The primary criticism of the current Army selection/classification system has been that it is not sufficiently linked to the job performance of the people it classifies. The explanation for the lack of linkage is compelling--credible measures of job performance do not exist. Current measures of job proficiency, Skill Qualification Tests, are designed and administered as diagnostic training tools rather than as indicators of successful job performance. As a result, current validity information on selection and classification is based on indicators of training success rather than job performance.

Not only does Project A propose to expand the criterion emphasis from training performance to job performance, but also to develop a model of job performance that encompasses more than some single dimension presumed to represent the job. The logical concomitant to this goal is that the factors that are defined as components of the job must be operationally defined and measurable. Job performance criteria being developed in Project A are of two major types. One type involves the measurement of overall performance as a soldier, that is, of constructs that apply to all Military Occupational Specialties (MOS) (Pulakos & Borman, 1985). The other type involves the measurement of MOS-specific performance, taking into account the differences as well as the commonalities in the requirements of different MOS.

MOS-specific measures are further divided into two classes. One class is that of behavior-based rating scales, which focus on identifying incidents that differentiate between successful and unsuccessful performance and grouping such incidents into rating dimensions (see Toquam, et al., 1985). The other class of MOS-specific measures, covered in this report, concerns task-based measures. This class involves comprehensive assessment, using hands-on, job knowledge, and rating measures, of performance on a set of tasks identified as best representing the major aspects of the MOS.

SPECIFIC OBJECTIVES

The specific objective of the work reported here was to develop reliable, valid, and economical task-based measures of first-tour job performance of enlisted personnel in nine MOS.

Hands-on tests are generally viewed as the most direct validation criteria, since they call for application of knowledge and demonstration of skill by eliciting behaviors that are equivalent, or nearly equivalent, to those required in the job setting. In a comparison of knowledge tests and performance tests in four Army occupational specialties, Vineberg and Taylor (1972) concluded that "where job performance relies almost solely on skill, job sample tests, or some other variety of performance measure, are essential" (p. 17). Consequently, the model of overall soldier effectiveness would not be complete without hands-on measurement of the specific job tasks that a soldier is expected to perform.

Although performance measurement specialists tend to favor the hands-on method of testing, the value of written tests is no less deserving of emphasis. Their greatest advantage lies in the economy of testing. With little equipment or personnel support, large groups of soldiers can be effectively tested in a short period of time. Furthermore, a knowledge test is a valid, and even preferable, mode of testing for tasks involving cognitive skills such as decision making, problem solving, and related applications of rules and principles. Vineberg and Taylor (1972) further concluded that "job knowledge tests can be appropriately substituted for job sample [performance] tests, when a job contains little or no skill component and when only knowledge required on the job is used in the test" (p. 19). Finally, knowledge tests are not subject to environmental and safety conditions that can constrain hands-on tests; thus greater breadth of task coverage can often be achieved with a knowledge test. The key to the utility of knowledge tests appears to be in the type of task or portion of the task they are used to evaluate. If used for the appropriate kind of behavior and linked methodically to knowledge-based task elements, knowledge tests have wide applicability and acceptable validity, and are exceptionally efficient.

For Project A, however, knowledge tests are included as criterion measures for more than reasons of economy. Although having the technical knowledge of how to perform tasks must certainly be a factor in demonstrating that knowledge, yet having the knowledge and being able to use it are not the same thing. Both hands-on tests and job knowledge tests can provide useful information about an individual's overall job performance. Each has limitations, however, and neither can be assumed in advance to represent the job incumbent's "true" performance. Intercorrelations between the two types of measures are of interest for what they tell us about the interchangeability of such measures. Are there tasks where such measures are highly interchangeable? If so, are both types of measures needed? Where correlations are low, are the measures tapping different aspects of the job? What implications do such findings have for the development of a composite measure of soldier performance? What conclusions can be drawn from the patterns of correlations across jobs with respect to the appropriate mixture of hands-on and job knowledge tests?

Both hands-on and knowledge tests have an acknowledged and established role as performance criterion instruments. The information that they provide, however, is incomplete in two respects (at least): Given that a soldier has the technical knowledge and skill, does he or she in fact perform the tasks consistently and correctly on the job? And, to what extent do recency and frequency of task performance help to explain test

results? These questions were addressed in two auxiliary instruments: the Task-Specific Performance Rating Scales, and the Job History Questionnaire.

The target population is Army enlisted soldiers in their first tour, covering approximately their first three years of duty. There are more than 250 MOS, which are generally equivalent to jobs, that soldiers may enter for their first tour; from these, nine MOS were selected for study (see Table 1). The nine MOS were chosen to provide maximum coverage of the total array of knowledge, ability, and skill requirements of Army jobs (Campbell, 1983; Rosse, Borman, Campbell, & Osborn, 1983). A brief description of the duties of first-tour soldiers in each of the MOS is contained in Appendix A.*

Table 1

Military Occupational Specialties (MOS)
Selected for Criterion Test Development

Batch A	
13B	Cannon Crewman
64C	Motor Transport Operator
71L	Administrative Specialist
95B	Military Police

Batch B	
11B	Infantryman
19E	Armor Crewman
31C	Single Channel Radio Operator
63B	Light Wheel Vehicle Mechanic
91A	Medical Specialist

*The appendixes to this report are contained in separate volumes (see abstract).

Chapter 2

DEVELOPMENT

FRAMEWORK FOR DEVELOPMENT

Design strategy for the MOS-specific tests involved selection in each MOS of approximately 30 tasks that accurately reflected the individual's job domain. The number of tasks was dictated mainly by the time allocated for testing, and while time required for testing would differ with the nature of the particular task, 30 tasks for each MOS seemed reasonable as a planning figure.

In each MOS, all 30 tasks would be tested in the knowledge (written) mode. Fifteen of the 30 tasks would be also tested in the hands-on mode. Job history data covering recency and frequency of performance would be collected for all 30 tasks, and task performance ratings for the 15 tasks tested hands-on.

OVERVIEW OF DEVELOPMENT PROCEDURES

The MOS-specific task tests and the auxiliary instruments were developed in two phases. Tests for four of the MOS (designated the Batch A MOS) were developed and field tested before the development of the tests in the remaining five MOS (Batch B) began. While the procedures were generally the same for Batch A and Batch B, some lessons learned from the Batch A development were applied to Batch B. All nine MOS required individual variations because of particular circumstances.

The general procedure comprised eight major activities:

1. Define task domain.
2. Collect subject matter expert (SME) judgments.
3. Analyze SME judgments.
4. Select tasks to be tested.
5. Assign tasks to test mode.
6. Construct hands-on and knowledge tests.
7. Conduct pilot tests and make revisions.
8. Construct auxiliary instruments.

These eight major activities are discussed in detail in the following section. It should be noted that the activities were performed independently among the nine MOS; only during actual test construction was any effort shared between any MOS, and then only on tasks selected for testing in more than one MOS.

PROCEDURES

1. Define Task Domain

The job definition of a first-tour (Skill Level 1) soldier was derived from MOS-specific and common task Soldier's Manuals (SM) and from Army Occupational Survey Program (AOSP) results. The SM reflect what the soldier is expected to be able to perform, according to Army doctrine. The AOSP provides data on what the soldier actually performs on the job and in training. The two are not necessarily conflicting but neither are they always congruent; neither source could be ignored.

These three primary resources, accessed for all MOS, are more fully described below.

- MOS-Specific Soldier's Manuals (SM). Each MOS Proponent, the agency responsible for prescribing MOS policy and doctrine, prepares and publishes a Soldier's Manual that lists and describes tasks, by Skill Level, that soldiers in the MOS are doctrinally responsible for knowing and performing. The number of tasks varied widely among the nine MOS, from a low of 17 Skill Level 1 (SL1) tasks to more than 130 SL1 tasks.

- Soldier's Manual of Common Tasks (SMCT) (FM 21-2, 3 October 1983).[1] The SMCT describes tasks that each soldier in the Army, regardless of his or her MOS, must be able to perform. The 1983 version contains 78 SL1 tasks[2] and "supersedes any common tasks appearing in MOS-specific Soldier's Manuals" (p. vii).

  The distinction between Common Tasks and MOS-Specific tasks generates some controversy. Neither one should be thought of as "more important" to the job; in fact, it is doctrinally incorrect to make any distinction at all, although this practice is widespread even among MOS proponents. Many Common Tasks would be in MOS-Specific Soldier's Manuals if their presence had not been superseded by inclusion in the SMCT. Indeed, many are

---

[1]For Batch A MOS, the version of FM 21-2 in effect during task selection was the 1 December 1982 edition, containing 71 tasks.

[2]Although by doctrine soldiers are responsible for all tasks in the SMCT, there are exceptions. Six of the tasks concern the protective mask and are divided into the M17 model and the M24/M25 model. A soldier would be responsible for only the type mask assigned, the M17 being more widely used.
Thus in application, soldiers are not responsible for all 78 tasks.

MOS crucial tasks, such as M16A1 rifle tasks for the Infantryman, first-aid tasks for the 91A Medical Specialist, and decontamination and equipment camouflage tasks for the 64C Motor Transport Operator, 31C Single Channel Radio Operator, and 19E Armor Crewman.[1]

- Army Occupational Survey Program (AOSP). The AOSP uses questionnaires prepared in conjunction with the MOS Proponent, with technical review provided by senior noncommissioned officers (NCO) with field experience in the MOS. The questionnaires present a list of tasks or part-tasks that usually include and expand the doctrinal tasks from the preceding two sources. The AOSP is administered periodically to soldiers in all skill levels of each MOS by the U.S. Army Soldier Support Center. Although the AOSP questionnaires make no distinction in the task listings regarding doctrinal skill level, the analysis of responses by means of the Comprehensive Occupational Data Analysis Program (CODAP) provides the number and percent of soldiers at each skill level who report that they perform each task. The number of tasks or activities in the surveys for the nine MOS of interest ranged from 487 to well over 800.

The above sources provided the main input to the individual MOS job domain definition. However, MOS Proponents were also contacted to determine whether other accepted task lists existed. While the additional tasks thus generated were not large in number, they were sometimes significant. For example, the pending introduction of new equipment, such as the M249 Squad Automatic Weapon to 11B (Infantryman) and 95B (Military Police), added tasks that had not yet appeared in the accessed documents.

The gathering of task lists was a preliminary activity and resulted in a large and not very orderly or meaningful accumulation of tasks, part tasks, steps, and activities. To bring some order to this accumulation, a six-step process was conducted for each MOS. This procedure included:

- Identify AOSP/CODAP activities performed at SL1.
- Group AOSP statements under SM tasks.
- Group AOSP-only statements into tasks.
- Consolidate domain (proponent review).
- Delete tasks that pertain only to restricted duty positions.
- Delete higher skill level and AOSP-only tasks with atypically low frequencies.

---

[1]Despite this philosophical admixture of all tasks, the project and this report continue to categorize tasks as "common" or "MOS-specific" primarily as an easy means of reference. However, treatment of all tasks was essentially the same regardless of their source.

All steps except the proponent review were performed by project staff who were familiar with the MOS and were experienced in job and task analysis.

Identify AOSP activities performed at SL1. AOSP/CODAP information of interest to this project was data regarding frequency of performance by SL1 personnel of each task or activity listed in the survey. The assumption for this step was that every activity included in an AOSP survey that had a non-zero response frequency among SL1 soldiers, after allowing for error in the survey, was performed at SL1. The procedure for estimating the error was to compute the average SL1 response frequency for each MOS survey and use that proportion to determine the boundaries of a confidence interval about zero. Tasks or activities with frequencies above the confidence interval boundary were considered to have non-zero frequencies and were retained for the next two steps; those below were dropped. For each MOS, about 25% of the tasks/activities on the AOSP were dropped by this application.

Group AOSP statements under SM tasks. An AOSP statement was referenced to an SM task if the statement duplicated the SM task or was subsumed under the SM task as a step or variation in conditions. The effort first tried to identify SL1 tasks (either MOS-specific or Common) with which the AOSP statement could be matched. If this could not be done, higher skill levels (HSL)--SL 2, 3, and 4--were successively reviewed and the AOSP statements matched with these SM tasks, if possible. Thus the grouping concentrated on matching AOSP statements with SM tasks wherever possible, even if doctrine did not specifically identify the activity as a SL1 responsibility. The resulting task list included all SL1 SM tasks (MOS-specific and Common) regardless of whether they had parallel or supporting AOSP statements, and all HSL tasks for which matching AOSP statements were found.

Group AOSP-only statements into tasks. Since some AOSP statements could not be matched with any SM task, or any subset of elements from an SM task, the next step was to edit these statements so that, although they were similar in format to the SM task statements, they were still a clear portrayal of additional task content not contained in the SM. In some cases an AOSP statement became a task statement by itself; in other cases, a new task statement was developed which could appropriately subsume several AOSP statements.

Consolidate domain (Proponent review). After the data were grouped, the result was a fairly orderly array of tasks for each MOS. With this task list it was possible to go to the proponent Army agency for each MOS to verify the tasks as being bona fide domain tasks. A minimum of three senior NCO or officers reviewed the list at each Proponent and tasks that were erroneously included in the domain were eliminated. While specific reasons for dropping tasks varied with each MOS, the general categories were:

- Non-Applicable Systems - In some MOS, emerging but overlapping equipment systems forced a decision either to concentrate on the emerging system or to try to cover all existing systems. In the case of MOS 19E, the armor crewman can potentially be assigned to an M60A3, M60A1, M48A5, or M551 tank. The M48A5 and M551 are virtually non-existent in the active duty inventory,

and by 1990 practically all of the M60A1 tanks are scheduled to be converted to M60A3 models.[1] Therefore, all equipment-specific tasks except those pertaining to the M60A3 were dropped. Likewise, concentrating on the A3 version resulted in dropping several nonapplicable machinegun tasks as well as infrared and searchlight tasks that do not apply to the M60A3.

- Eliminated by Doctrine - Advancing and active doctrine sometimes results in a lag between what is in the SM and what is emerging as doctrine. For example, a DA Message (March 1984) citing the Geneva Convention eliminated 12 tasks involving certain offensive weapons from the 91A (Medical Specialist) domain list. More often there is a gap between the infrequently generated AOSP task list and what is current doctrine. For example, dropping Equipment Serviceability Criteria Inspection requirements, changes in first aid procedures under nuclear-biological-chemical conditions, and consolidation of urban terrain combat tasks all resulted in tasks being eliminated from various domains.

- Collective Tasks - Some tasks were included that are actually collective tasks performed by crews/squads, platoons, or even companies/batteries. In all cases these collective tasks encompassed individual tasks that were covered elsewhere.

- Combined Systems - In the 13B (Cannon Crewman) MOS, an SL1 incumbent can be assigned as a crewman on one of six howitzers. However, the SM lists many howitzer-specific tasks six times (once for each howitzer). Therefore, these tasks were collapsed into a single task. The rationale was that a soldier will be responsible for performing the task on only one type of gun, not all six. If a task thus collapsed was selected for testing, six versions of the test might be necessary, but at this point a single listing of such tasks was preferred.

- Reserve Component - While for most units there are no discriminable differences between active duty and Reserve Component organizations, this is not true for all MOS. Many of the differences are equipment-specific, and others are the result of mission differences. Because the project was concerned only with active duty performance, tasks that applied only to Reserve Component incumbents were dropped.

The full consolidated domain list of tasks, with supporting AOSP statements for each MOS, is contained in Appendix B.

Delete tasks that pertain only to restricted duty positions. The SM for most MOS contain tasks for individual duty positions within the MOS. For example, an 11B Infantryman can be a Radio-Telephone Operator, Machine-gunner, Grenadier, Scout, Driver, or Dragon Gunner; the 64C Motor Transport

---

[1]Because some M60A1 units were encountered during field testing, some test modification was required during this phase.

Operator can be a Dispatcher; the 95B Military Policeman can be a Security Guard. For most duty positions, incumbents move freely in and out of the position, the performance of the duty position tasks being dependent on whether or not the soldier is assigned to the position. Other positions are more permanent.

To deal with the problem of duty position variations, Restricted Duty Positions were operationally defined as those for which the award of an Additional Skill Identifier or Special Skill Identifier and at least one week of specialized training were required. The tasks specific to duty positions that met this criterion were dropped. Only five duty positions in two MOS were affected. In 13B (Cannon Crewman) these were the M198 Artillery Mechanics, Assemblers of 155mm Atomic Projectiles, Assemblers of the 8-inch Atomic Projectile, and Nuclear Security Guards. In 71L (Administrative Specialist), the Postal Clerk met this criterion. All other duty position tasks were retained under the assumption that an incumbent could be expected to fill the duty position at any time.

Delete HSL and AOSP-only tasks with atypically low frequencies. The domain task lists that were generated from the initial collection and consolidation of data reflected a fairly complete and broad definition of what the SL1 soldier could be expected to encounter on the job, because of either job requirements or doctrinal policy. However, the domain listing also contained many tasks, especially HSL tasks, that were performed so infrequently that their presence was not representative of SL1 MOS expectations. Since the domain listings were still quite large and unwieldy for making criticality and clustering decisions, the domains were refined by deleting those HSL and AOSP-only tasks with low frequencies.

The first step in this process was to translate AOSP/CODAP frequencies into task frequencies. Performance frequencies were available from the CODAP analysis on all AOSP statements. These frequencies are not reliable, in the judgment of some proponent reviewers, who believe that respondents tend not to discriminate in their responses or to accurately define the task they are responding to. The reviewers commented that SL1 soldiers never perform certain of the supervisory tasks, which are the duties of senior NCO; the fact that a few SL1 soldiers reported having done so lead them to consider all of the survey data suspect. Nonetheless, the CODAP results are the best available documentation of how widespread the performance of the task is. Additionally, although the frequency of performance may be questioned as an absolute percentage, it is probably highly accurate when used for comparison among tasks within the survey. However, a problem with using the CODAP frequencies was that not all AOSP statements corresponded with task statements. In many cases, the AOSP statements represented steps within the tasks; in other cases, several AOSP statements represented various conditions under which a single SM task could be performed.

To resolve these conflicts and to arrive at a representative frequency for each task, the algorithm shown in Figure 1 was developed. Generally, when AOSP and task statements matched, the CODAP frequency for the matching statement was applied to the task. If there was no match, the most frequent step or condition was used as the basis for the task frequency. However, in some cases, frequencies were aggregated to account for equipment differences.

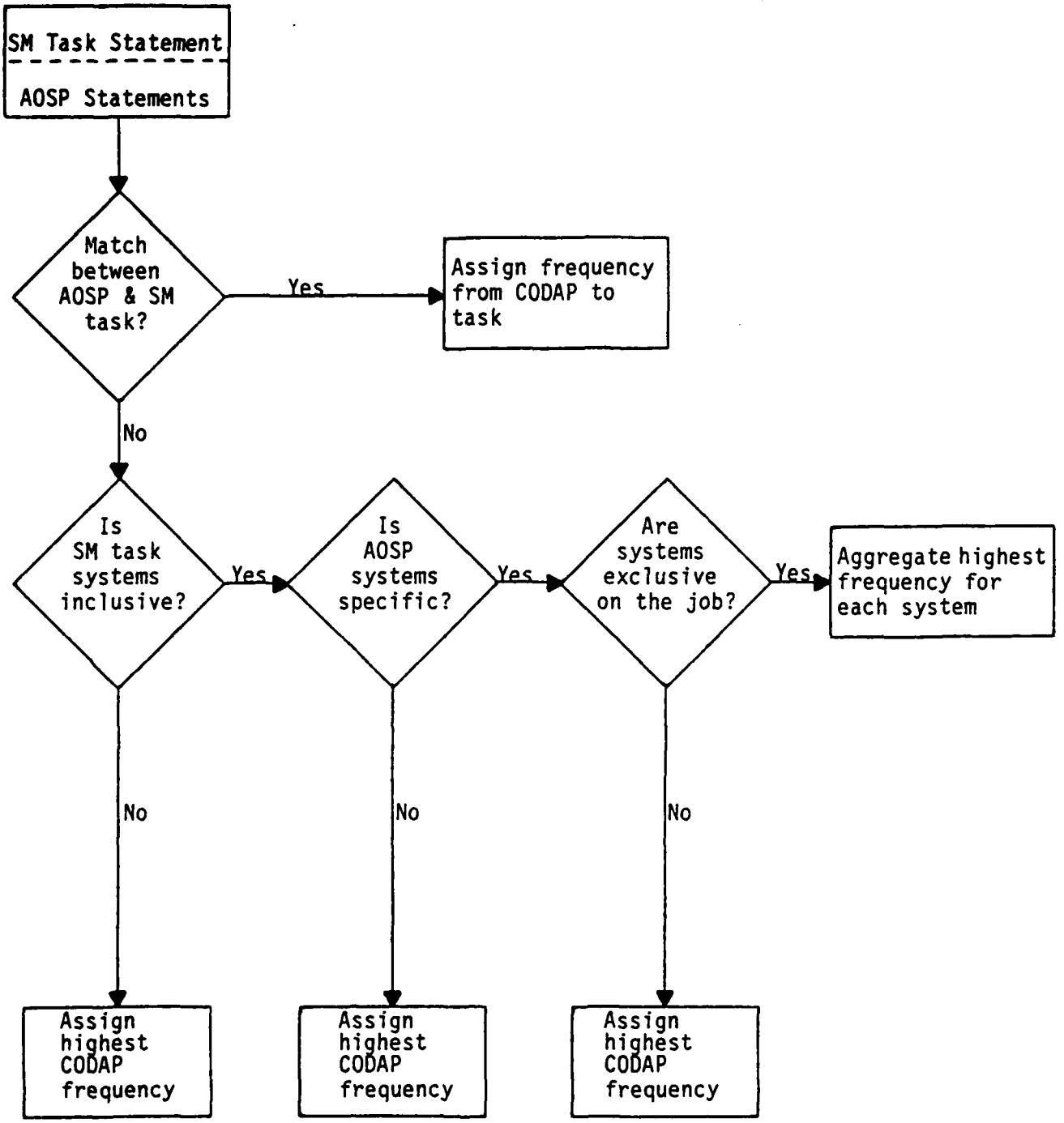

Figure 1. Method for assigning performance frequencies to tasks

The general approach for identifying low-frequency tasks was to compare frequency distributions of the SL1 tasks (MOS and Common) with the HSL and AOSP-only tasks. HSL and AOSP-only tasks were then eliminated until the two groups were not significantly different with respect to location, dispersion, and form. A four-step procedure identified the atypically infrequent tasks to be eliminated:

- List the response frequencies of SL1 tasks from the AOSP/CODAP.
- List the response frequencies of HSL/AOSP-only tasks.
- Test groups (lists) for difference, using Mann-Whitney U test.
- If groups were different, and the HSL/AOSP-only group had tasks with lower response frequencies (which they did in all cases), eliminate those low frequency tasks until group differences were not significant at .01 level.

The result was a final task list for each MOS. It included all SL1 MOS and Common Tasks with non-zero frequencies (or no AOSP/CODAP frequency), and HSL/AOSP-only tasks performed by SL1 soldiers. Table 2 shows the reduction of the task list during each phase and the reasons for the reduction, by MOS. The nine final task lists are contained (with data from the SME judgments, described below) in Appendix C.

2. Collect SME Judgments

After the MOS domains were refined, every domain contained more than 100 tasks. To select 30 tasks for each MOS that would represent most of the domain, that would include the most critical tasks for the MOS, and that would have a sufficient range of performance difficulty to permit some discrimination among soldiers, additional information was needed.

MOS Proponent agencies were asked to provide SME to render judgments regarding the tasks on the task list. Requirements for SME were that they be NCO in the grade E-6 or above (i.e., second or third tour) or officers in the grade O-3 (Captain) or above, and either hold or have recently held the MOS being reviewed. Recent field experience--that is, assignment to a unit supervising SL1 personnel in the MOS--was an additional requirement. For the Batch A MOS, 15 SME in each MOS were requested. For the Batch B MOS, some modifications were made in the review process (described below) and 30 SME in each of these MOS were requested. Collection of SME data required approximately one day for each MOS. Three types of judgments were obtained from the SME:

- Task clustering
- Task importance
- Task performance

The number of SME obtained for each MOS and samples of all instructions provided to SME are contained in Appendix D.

Table 2

Effects of Domain Definition on Task Lists

AOSP Review	13B	64C	71L	95B	11B	19E	31C	63B	91A
AOSP Statements	669	677	822	546	822	609	656	633	685
Deleted - Zero Frequency	67	169	329	197	188	103	134	84	267
Deleted by SME	--	--	58	--	--	--	--	195	61
AOSP Statements Used	602	508	435	369	634	506	522	354	357

Domain Consolidation	13B	64C	71L	95B	11B	19E	31C	63B	91A
Tasks in MOS	378	166	203	304	357	338	267	188	251
Nonapplicable Systems	-	-	-	-	-	50	-	-	-
Eliminated By Doctrine	23	-	-	-	16	14	97	10	12
Collective Tasks	25	-	-	-	5	-	-	-	-
Combined Systems	57	-	-	-	-	-	-	-	-
Reserve Component Tasks	-	-	-	-	15	-	-	-	-
Tasks in Domain	273	166	203	304	321	274	170	178	239

Domain Reduction	13B	64C	71L	95B	11B	19E	31C	63B	91A
Tasks in Domain	273	166	203	304	321	274	170	178	239
Restricted Duty Position	44	-	42	-	-	-	-	-	-
Preliminary Sort	-	-	-	176	-	-	-	-	-
Low Frequency-HSL/AOSP Only	53	47	-	-	90	39	-	-	-
Domain Tasks For SME Judgments	177	119	161	128	231	235	170	178	239

Task clustering. Each task title was listed on a 3"x5" index card along with a brief (3 or 4 line) description of what the task entailed. Each task was code numbered (see Appendix D for sample). For Batch A MOS, the cards were randomly shuffled to intermix Common and MOS tasks and to break up normally occurring sequences of tasks. SME were told to sort the tasks into groups so that all the tasks in each group were alike, and each group was different from the other groups. For the Batch B MOS, common tasks were grouped for the SME, based on the clustering derived from the Batch A data. SME were permitted to add to or break up the groups as they saw fit.

Task importance. These judgments required the SME to consider importance of tasks in a job setting. Since the environment of the job setting can affect the outcome, an attempt was made to standardize the SME raters' point of view by means of a context scenario. For the Batch A MOS, all SME were given a European scenario which specified a high state of training and strategic readiness but was short of involving actual conflict. This so-called "neutral" scenario is shown in Appendix D for MOS 71L (Administrative Specialist). Very slight modifications were made in the terminology for the other MOS to make them specific to that MOS.

Following collection of Batch A MOS data, there was concern about the effect of scenario on SME judgments. Although a single context scenario, as used in Batch A, was presumed to standardize raters' views, rethinking the issue led to the decision to elicit judgments of task importance under a range of hypothetical conditions. As a result, for Batch B MOS three scenarios were used. The neutral scenario (now called an "Increasing Tension" scenario), identical to that used in Batch A, was retained. A "Training" scenario specifying a stateside environment was developed, as well as a "Combat" scenario (European, non-nuclear). These latter two scenarios are shown in Appendix D for MOS 95B (Military Police) and 19E (Armor Crewman). Again, minor terminology modifications were made for the other MOS. The SME for each Batch B MOS were divided into three groups and each group was given a different scenario as a basis for judgments. For the 63B (Light Wheel Vehicle Mechanic), however, only 11 SME participated; each SME rated task importance three times, using each of the three scenarios in counterbalanced order.

For these judgments, SME used the scenarios and were told to rate the importance of the task in performing the MOS job in support of the unit mission. Slightly different approaches were used in the Batch A and Batch B data collection. For Batch A MOS, the SME were given the tasks on individual cards identical to those used in task clustering and told to rank them from Most Important to Least Important. In this ranking no two tasks could be ranked the same. For Batch B MOS, a scale ranking was used. Instead of the cards, SME were provided a list of the tasks and their descriptions and asked to rate them on a 7-point scale from "1 = Not at all important for unit success" to "7 = Absolutely essential for unit success." These SME could, of course, give different tasks the same rank.

Task performance. To obtain an indication of expected task performance distribution, SME were asked to sort soldiers into five categories based on how they would expect a typical group of 10 SL1 soldiers to be able to perform on each task. The instructions provided the SME for making this

rating and a description of the five performance levels are presented in Appendix D. Scenarios were not used in this judgment.

3. Analyze SME Judgments

The judgment data were analyzed and the following products were obtained:

- Cluster Membership - For each task, the cluster to which the task was assigned, based on a factor analysis of a cross-products matrix derived from the SME's task similarity clusterings.

- Similarity - For each task, the mean number of SME who placed the task in a cluster with each other task. This information was used primarily in making decisions about borderline tasks in the cluster analysis. It was included in the subsequent data presentation for Batch A tasks, but because it had little meaning after final clustering it was dropped from the information provided for Batch B tasks.

- Importance - The rank of each task, averaged across SME. For Batch A MOS, a single importance score was obtained. For Batch B MOS, three separate importance rankings were obtained, one for each of the three scenarios. In all cases, a low numerical rank indicated that the task was judged Very Important.

- Difficulty Mean - The mean of the distribution of 10 hypothetical soldiers to the five difficulty levels on each task, averaged across SME. A high number indicated that the task was judged as very difficult.

- Performance Variability - The standard deviation of the distribution of 10 hypothetical soldiers to the five difficulty levels on each task, averaged across SME. Low numbers indicate tasks where little variability could be expected.

Printouts were produced which listed tasks by cluster, and within clusters, ordered by Importance rank. The tables also included Difficulty Mean, Performance Variability, and Frequency (if available) for each task. The frequencies listed were taken from the AOSP/CODAP; for Batch B, these were supplemented with estimates obtained during the SME domain review. For the Batch B MOS, which presented ranking under three scenarios, the sequence within clusters was determined by the Combat Scenario ranks. These tables are contained in Appendix C.

4. Select Tasks to Be Tested

Based on previous experience, project staff estimated that approximately 30 tasks per MOS could be tested within the time and resource constraints of troop support organizations. While the methods used for selection were similar for Batch A and Batch B MOS, they varied enough between batches that they are outlined separately.

Batch A test task selection. From five to nine project personnel participated in the selection process on each MOS; all were experienced in military testing, and had participated in the project process to date. For each MOS the individual who had prime responsibility for the particular MOS, and who had by that time developed the greatest familiarity with the MOS, was one of the participants.

The task selection panel was provided the data summaries described above (see Appendix C) and asked to select 35 tasks to represent each MOS. (In anticipation of internal and external reviews, each MOS was oversampled by five tasks.) No strict rules were imposed on the analysts in making their selections, although they were told that high importance, difficulty, variability, and frequency were desirable, and that each cluster should be sampled. Thus, selection utilized a judgment-based approach, without rigid parametric constraints, that allowed idiosyncratic judgments of task analysts to enter the selection process.

In the first phase of the process, each analyst made his or her selections. The results were analyzed with the objective of capturing each individual's policy for use of the data. For each analyst, task selections were first regressed on the task characteristics data to identify individual selection policies. The equations were then applied to the task characteristics data to provide a prediction of the task selections each would have made if their selections were completely consistent with their general tendency, as represented by the linear model.

In the second phase, analysts were given their original task selections and the selections predicted by their regression-captured policies. They were directed to review and justify discrepancies between their observed and predicted selections. Analysts independently either modified their selections or justified their original selections. Rationale for intentional discrepancies was identified and the regression equations adjusted.

The last phase of the analysts' selection procedure was a Delphi-type negotiation among analysts to converge their respective choices into 35 tasks for each MOS. (Instructions to analysts are shown in Appendix E.) The choices and rationale provided by analysts in the preceding phase were distributed to all analysts, and each made a decision to retain or adjust his or her decisions, taking into account the opinions of others. Decisions and revisions were collected, collated, and redistributed as needed until near consensus was reached. For all MOS, three iterations were necessary. For MOS 13B and 64C, full consensus on 35 tasks was achieved in face-to-face discussions with all participants; for MOS 71L and 95B, consensus was achieved among four analysts, and their list of 35 tasks was communicated to the other group of analysts, who used the list as input in arriving at a consensus. For all MOS, the analysts designated 30 tasks as high priority for testing, and five tasks as alternate selections.

The resulting task selection was mailed to each Proponent; a briefing by Project A staff was provided if requested. A Proponent representative then coordinated a review of the list by proponent personnel designated as having the appropriate qualifications. After some minor Proponent-recommended adjustments, the final 30 tasks were selected. These are listed in Appendix F for the Batch A MOS.

Batch B test task selection. Based on experiences with Batch A selection, two major modifications were introduced in the selection process for Batch B. First, Proponent representatives were more actively involved in the selection process. Second, the decision was made to drop the regression analysis for Batch B selection. Experience with Batch A indicated that use of the regression analyses for policy-capturing would be prohibitively complex and time consuming when non-project participants were integrated into the process. More important, the Batch A experience showed analysts' selections to be non-linear; they qualified their selections on the basis of knowledge of the MOS or the tasks, information not represented in the data provided. And while analysts used the data extensively, they used it in a non-linear combination that often differed with each cluster. Thus, to summarize the non-linear judgments using a linear analysis model provided a relatively poor description of the analysts' processes.

The panel for Batch B selection consisted of 5-9 members of the project staff as in Batch A, combined with six military personnel (NCO and officers) from each MOS. These six were in the grade of E-6 or higher with recent field experience, and were selected to provide minority and gender (for applicable MOS) representation to the task selection process. This latter factor was introduced to minimize the possibility of subsequent criticism of cultural/gender bias in the tasks used for testing.

The materials provided the selection panel were the same variables generated by the SME judgments (less the Similarity data provided for Batch A). Again, no strict rules were imposed. However, panel members were provided a target number of tasks to be selected from each cluster, calculated in proportion to the number of tasks in each cluster, with a total of 35 tasks to be selected. A second adjustment prescribed a minimum of two tasks per cluster to permit estimation of the correlation among tasks in the cluster.

The initial selection phases were performed independently by panel personnel (see Appendix E for the instructions provided the panel members for the initial session). For Session 2 each panel member was provided a composite record of the choices made by the panel and asked to independently select again; this time they were asked to write a brief justification for their selections. For Session 3 panel members were provided the latest composite selections along with composite selection reasons and asked again to independently select tasks. In the last selection session one or two project staff members met with the military representatives (the requirement to be at the Proponent site precluded participation of all panel members in this fourth session). Members were provided the latest selections and in a face-to-face meeting discussed and chose the final 30 tasks. (It should be noted that for most MOS, consensus had been reached on 70-80% of the tasks before this fourth session.) The tasks selected for Batch B MOS are listed in Appendix F.

5. Assign Tasks to Test Mode

The initial development effort required that knowledge tests be developed for all 30 tasks, and hands-on tests for 15 of these tasks. The procedure for determining the suitability of a task for the hands-on mode of

testing was based on recognition that the types of tasks vary in suitability for measurement in a knowledge mode, while the total number of tasks (30) exceeds the hands-on resources. The considerations involved in selection for hands-on testing were:

- Fifteen soldiers must complete all hands-on tests in four hours. No single test may take more than 20 minutes unless several soldiers can be tested at a time.

- Some degree of simulation is necessary in performance testing of many tasks since realistic task conditions may endanger people or equipment or may vary uncontrollably, preventing standardization.

- Scorer support would be limited to eight NCO scorers.

- All tests for an MOS would be tested in the same general location. The hands-on test site must be within walking distance of the other test activities. Terrain and site requirements therefore must be minimal.

- Equipment requirements must be kept within reason if units were to support the requirement. Additionally, modification or alteration of equipment for testing could not threaten the operational status of equipment provided.

- The test had to be administrable in a number of CONUS installations and Europe. Uniform testing would not be possible for tasks affected by local SOP.

- A hands-on test is much preferred for tasks that entail:
  - Physical strength or skilled psychomotor performance.
  - Performance within a time limit.
  - Many procedural steps.
  - Steps that are uncued in their normal sequence.

Based on these constraints, project test experts prepared an anticipated approach for hands-on testing for each of the 30 tasks in each MOS. The anticipated approach for each task included the following information:

- Equipment/Location - What the Army must provide to support the test.

- Simulation - What compromises will be necessary in translating job conditions into test conditions.

- Scope - What the soldier will do.

- Job Aids - Written material that the soldier will be allowed to use on the test.

- Steps or Subtasks Covered - Based on task analysis information.

- Estimated Time to Test One Soldier - Including preparation time between soldiers.

This worksheet, along with the task analysis data or SM extract for the task, was provided to five project analysts. Working independently, each analyst first reviewed the testing approach, and modified it as he/she deemed necessary. The analyst then worked with the test approach and supporting materials to assign points to each task to indicate hands-on test suitability, using the following three areas of consideration:

- Skill Requirements - Analysts determined a numerical value for skill requirements based on the number of steps requiring physical strength, control, or coordination. They also considered whether a step, although not inherently skilled, should be considered skilled if a doctrinal time limit was imposed. Finally, they determined whether steps that singly were not skilled but that must be performed together (integrated) to meet task requirements should collectively be considered skilled. Each skilled step was counted as one point, and each integrated set of steps was also counted as one point.

- Omission Value - This rating considered the likelihood that a soldier would omit a step on the job. For a step to have any omission value, three conditions were required:
  - A soldier must be able to complete the procedure (albeit incorrectly) without performing the step.
  - The step must be required.
  - Nothing in the test situation must cue the soldier to do the step.

  Each omission step received a numerical rating of one.

- Time Value - There were two levels of relevance for time. The clearest was where doctrine (usually the SM) specified a time limit for task performance; this was awarded a numerical value of two. The second was where no doctrinal time limit has been established but where performance could be inferred from time data, that is, where difference in time would be a reliable indication of task proficiency; this was awarded a numerical value of one.

Following the individual ratings, analysts met in group discussions and proceeded, task by task, to resolve differences until a consensus was reached and a single numerical score was assigned to each task. The tasks were then rank-ordered by score and a final feasibility check was conducted to insure that the top 15 rated tasks fell within the 4-hour time limit. These tasks were then earmarked for hands-on development.

One modification was applied to the hands-on selection process during test development. Common tasks were restricted to a single MOS, that is, no two MOS, in Batch A or Batch B or across Batches, were to use the same task in the hands-on tests. Instead, next-rated tasks were selected for development. The rationale was that hands-on data gathered in one MOS would suffice to apply to the task for all MOS, and that expanding the original

pool of hands-on tasks would provide needed flexibility if hands-on task adjustments later proved necessary. Exceptions were made in Batch B for tasks that could be expected to be performed substantially differently between MOS, such as "Perform Operator Maintenance on M16A1 Rifle" as performed by the 71L (Administrative Specialist) and 11B (Infantryman). Exceptions were also made when the set of 30 tasks selected for an MOS had fewer than 15 "new" tasks (not previously tested hands-on) that could be tested within the available four hours. Thus, for example, "Apply Field or Pressure Dressing" was tested hands-on in the 63B (Light Wheel Vehicle Mechanic) test, as well as in the 71L (Administrative Specialist) test.

6. Construct Hands-On and Knowledge Tests

For both hands-on and knowledge tests, the primary source of information was task analysis data. Task analyses were derived from the Soldier's Manuals, Technical Manuals (TM), and other supporting Army publications, as well as SME input and direct task observation, where necessary. Much of the development effort involved the individuals working on both types of test for the same tasks. Otherwise, the actual development of the two tests for tasks was independent; each task test was developed with an eye to taking full advantage of the capabilities of the test mode to achieve valid and reliable measurement of as much of the task as possible.

Hands-on test development. Hands-on test development followed a procedure designed to maximize inherent validity and enhance scorer reliability during test administration. The model for developing hands-on tests emphasized four areas:

- Determine test conditions.
- List performance measures.
- State examinee instructions.
- Develop scorer instructions.

These four activities are not separate or sequential actions. Each step in an activity interacts with the other activities and all must be integrated to make a complete test.

- Determine test conditions. Test conditions are what the soldier experiences during the test. They are designed to maximize the standardization of the test between test sites and among soldiers at the same test site. They are purposely restrictive yet must not be unnecessarily so. Test conditions are determined for the test environment, equipment, location, and task limits.

- List performance measures. The performance measures are the substance of the test--the behaviors to be rated GO/NO-GO by the scorer. Performance measures are defined as either product or process depending on what the scorer is directed to observe to

score the behavior.[1] Performance measures must adhere to the following principles:

- Describe observable behavior only.
- State a single pass/fail behavior.
- Contain only necessary actions.
- Contain a standard (how much or how well).
- State an error tolerance limit if variation in behavior is permissible.
- Include a scored time limit if, and only if, the task or step is doctrinally time-constrained.[2]
- Include a sequence requirement if, and only if, sequence is doctrinally required.

Performance measures are not designed to describe each action nor to score each behavior of the soldier, but to concentrate on those areas that best measure task behavior. They must seek a balance between comprehensiveness and allowing the scorer to concentrate on watching the soldier perform. Wordiness and complexity must be avoided if accurate and consistent scoring is to result.

- State examinee instructions. The examinee must be told what to do when he or she arrives at the test station. Most of the time the examinee can simply be told the name of the task which must be performed. However, when task limits have been modified for test purposes, this information must be conveyed to the examinee. The instructions must be kept very short and very simple; any information not absolutely essential to performance must be excluded. Examinee instructions are read verbatim to the soldier by the scorer and may be repeated at any time. However, these written instructions are the only verbal communication the scorer is allowed to have with the soldier during the test.

---

[1]It should be noted that tasks can also be scored process or product, but for this project process measures were also used even when a task product was obtained. For example, the task "Determine Azimuth With A Compass" can be scored solely by evaluating the compass reading obtained--the outcome. It can be evaluated by scoring the steps that should be done to obtain a correct compass reading--zeroing, holding the instrument, sighting. Since this project was interested in part-task performance (if a soldier could not perform the entire task), process measures were used throughout in addition to product measures where appropriate.

[2]During test administration, soldiers would be allowed to continue with the task even though they exceeded doctrinal time limits as long as they were making progress on task performance. They would, of course, be scored NO-GO on the time criterion.

- Develop scorer instructions. These instructions tell the scorer how to set up, administer, and score the test. They cover both usual and unusual situations, and insure standardized administration and scoring. The instructions include step-by-step procedures for setting up the station, setting up equipment, and restoring equipment and conditions for subsequent administration, as well as specific scoring techniques that are not included in the performance measures. It is essential that the scorer become totally familiar with all scorer instructions before the first test is administered.

Knowledge test development. Knowledge tests, to be valid and reliable, must concentrate on those performance items for which they are suitable. Knowledge tests do not necessarily mirror the measures in a performance test, nor should they seek to elicit the "why" of performance at the expense of "how to."

The format of knowledge tests is dictated to some extent by their proposed use. For example, a free-response format can be used to test knowledge of a task sequence, but such formats demand more of the soldier's literacy skills and are more difficult to score reliably. The multiple-choice format is easier to score and is familiar to most soldiers. However, it is more difficult to develop because of inherent cueing, particularly between items, and the need to develop likely and plausible but clearly wrong alternatives. Because of the quantity of data to be gathered in the project, machine scoring was essential. Therefore a multiple choice format with 2-5 choices and a single correct response was selected.

Knowledge tests, unlike hands-on tests, can cover the task under varying conditions and circumstances and do not require adjustment in task limits to meet time constraints. But knowledge tasks are not without test administration constraints dictated by the number of tasks to be tested and the time available for testing. For the project, all tasks selected (approximately 30 per MOS) would be tested in the knowledge mode. Four hours were allocated to the knowledge testing block for the field trials, to be reduced to two hours for Concurrent Validation testing. Allowing an average of slightly less than one minute to read and answer each item dictated an average of about nine items per task. Thus task coverage had to be restricted by the same type time constraints that affected hands-on testing.

Knowledge test development is based on the same information available for hands-on development, namely the task analysis. But there are some additional requirements not directly encountered in hands-on test development, and knowledge tests rely more on task input from SME beyond what is contained in the task analysis. The three distinct characteristics of multiple choice performance knowledge test items are that they (a) are performance or performance-based measures, (b) identify performance errors, and (c) present likely alternatives.

- Are performance or performance-based. Knowledge test items are considered either performance or performance-based, depending on the actions required of the examinee. Where those activities are the same as on the job, the test is a performance

test even though given in the written or knowledge mode. For example, the tasks "Determine Distance On A Map" and "Authenticate Using A CEOI" require the same actions in a knowledge test as on a hands-on test, and are examples of knowledge performance tests. After the examinee determines the distance, or reads the CEOI to find the authentication, he or she refers back to the answer alternatives to match the obtained answer.

Most tasks, of course, cannot elicit full job-like behavior in the knowledge mode and therefore must be tested using performance-based items. These items require the examinee to select an answer describing how something should be done. For example, a task such as "Move Under Direct Fire" would ask the examinee how to cross an open space, or how to carry a rifle in a low crawl position. The test still focuses on how the task is done.

A prevalent pitfall in developing performance knowledge tests is a tendency to cover information about why a step or action is done or to rely on technical questions about the task or equipment. Just as in the hands-on tests, the objective of the knowledge test is to measure the soldier's ability to perform a task. The test must concentrate on the application of knowledge to perform a task, and the knowledge or recall required must not exceed what is required of a soldier actually performing the task. Because of this performance requirement, knowledge tests must present job-relevant stimuli as much as possible, and the liberal use of quality illustrations is essential.

- Identify performance errors. Performance-based knowledge tests must focus on what soldiers do when they fail to perform the task or steps in the task correctly. This information must be obtained by experience with actual soldiers. The approach to test development used posits four causes of low performance proficiency and each step in the task must be analyzed for these four causes. Sometimes more than one will apply, but it is important to first identify where task failure most frequently occurs. The four causes are:
  - Does not know where to perform. These are location problems, most often associated with components of equipment.
  - Does not know when to perform. Usually associated with sequence problems of multi-step tasks.
  - Does not know what the product is. Involves cue or condition recognition and is usually tied in with an action or reaction.
  - Does not know how to perform the procedure. Involves execution of a step or series of steps.

- Present likely alternatives. The easiest alternative to write is the correct one. The incorrect alternatives present the most challenge for they must not only be incorrect but also be likely. The approach focuses on identifying what it is that soldiers do wrong when they perform a step--that is, if they do not perform

the step correctly, what is it that they do perform. This becomes the basis for the other alternatives. Some alternatives are, of course, more likely than others but all must be possible. Incorrect alternatives were limited to four by format design, but in some cases only one or two "real world" alternatives were possible and these were all that were listed. Finally, incorrect alternatives must be, in fact, incorrect; the correct alternative may not be merely preferred or "better."

Knowledge tests were constructed by project personnel with experience in test item construction and expertise in the MOS/task being tested. Developers based task coverage on knowledge of causes of task failure as outlined above. Tests were reviewed internally by a panel of test experts to insure consistency between individual developers. The following general guidelines were followed in construction:

- Stem length was usually restricted to two lines. Where necessary, a Situation was used if it could be applied to two or more items. The overall effort was to minimize the reading skills necessary to take the test.

- Item stems were designed so that the item could be answered based on the stem alone, that is, without reference to the alternatives.

- Illustrations were used where they duplicated job cues. Where necessary, illustrations were also used as alternatives as well as to provide a job-related reference. All illustrations were drawings. Test developers prepared the basic illustration, using a photograph or TM illustration. Final visuals were prepared by professional staff artists and draftpersons and checked by the developer before final printing.

- Each task tested was a separate entity clearly identified by task title and separate from other tested tasks.

- Extracts were prepared for tasks that allowed or required use of publications on the job. If the publication was lengthy, as, for example, tables of vehicle maintenance checks, the extract was provided as a separate hand-out. Brief extracts, of one page or less, were appended to the test. Materials for performance knowledge tests, such as maps, protractors, and scratch paper, were also provided.

- Test items within a task were arranged in the sequence that they would normally occur when the soldier performed the task.

- Completed tests were checked for inter-item cueing.

- All correct alternatives were authenticated as correct by a citable reference.

In four of the nine MOS, some of the tasks that incumbents perform are affected by the equipment to which they are assigned. Not all tasks are affected; for each of the four MOS a substantial number of tasks are "core" tasks and are the same regardless of equipment. However, sufficient differences exist in some of the other tasks to require specific treatment. The MOS affected were:

- 13B (Cannon Crewman) - Incumbents can be assigned to either the M109, M110, M198, or M102 howitzer.
- 19E (Armor Crewman) - Incumbents can be assigned to either the M60A1 or M60A3 tank.
- 11B (Infantryman) - Incumbents can be assigned to units that are either non-mechanized infantry or mechanized infantry.
- 95B (Military Police) - Most male incumbents carry a .45 caliber pistol, while some males and all females carry a .38 caliber pistol.

For these MOS it was necessary to develop separate tracked versions of some tests covering the specific items of equipment.

7. Conduct Pilot Tests and Make Revisions

Following construction of the tests, arrangements were made through the Proponent for troop support to pilot test the hands-on and knowledge tests. This procedure was conducted by the test developer and involved the support of four NCO scorers/SME, five MOS incumbents in SL1, and the equipment dictated by the hands-on test.

Pilot of hands-on tests. The following activities were performed:

- Test review - The four scorers independently reviewed the Instructions to the Scorer and scoresheets. The developer noted comments or questions that could be clarified by changes or additions to the materials.
- Test set up - One of the scorers set up the test as directed in the prepared instructions. The developer noted deficiencies or changes in the instructions.
- Scoring - One of the incumbents performed the test while the four scorers scored the test independently. After the test, all four scoresheets were compared. Discrepancies in scoring were discussed and the reasons ascertained. Some scorer discrepancies were the result of a scorer's physical position relative to the incumbent, but many required changing a performance measure or Instructions to Scorer, or even changes in the test or performance procedure itself. If possible, these changes were made before the next incumbent was tested. Normally, variations in incumbent performances occur naturally

but to insure variation the developer can cue incorrect performance unknown to the scorers. Testing continued with each incumbent, followed by the scoresheet review and revision. The incumbent was included in the review process to assist in determining how he or she actually performed.

- Examinee debriefing - Incumbents were interviewed to determine whether the instructions provided them adequate guidance on what they were expected to perform.

- Time data collection - Performance times were kept on all incumbents, as well as station and test set-up times.

Based on the pilot test information, a final version of each hands-on test was prepared. These tests are contained in Appendix G (limited distribution).

Pilot of knowledge tests. The knowledge tests were piloted at the same time that the pilot tests of the hands-on measures were conducted, utilizing the same four NCO hands-on scorers and the five MOS incumbents. The procedure was different for the two groups.

- NCO SME. The test developer went through each test item by item with all four NCO simultaneously. The specific questions addressed were as follows:
  - Would the SL1 soldier be expected to perform this step, make this decision, or possess this knowledge in the performance of this task on the job?
  - Is the keyed alternative correct?
  - Are the incorrect alternatives incorrect?
  - Is there anything in local or unit SOP that would affect the way this task item is performed?
  - Are the illustrations clear, necessary, and sufficient?
  - Is there any aspect of this task that is not covered in the test which should be covered?

- Incumbents. The five incumbents took the test as actual examinees. They were briefed as to the purpose of the pilot test and told to attempt to answer all items. The tests were administered by task and the time to complete each task test was recorded individually. After each task test the incumbents were debriefed. The following questions were addressed:
  - Were there any items where you did not understand what you were supposed to do or answer?
  - Were there any words that you did not know the meaning of?
  - Were there any illustrations that you did not understand?
  - (Item by item) This is what is supposed to be the correct answer for Item __. Regardless of how you answered it, do you agree or disagree that that choice should be correct?

The tests were revised on the basis of SME and incumbent inputs. Average completion time for incumbents was calculated for each task. On the basis of these times, the tests were divided into four booklets of approximately equal lengths for Batch A MOS; for Batch B, tests were divided into two booklets. By dividing tests into several booklets and varying the order of administration among groups of soldiers, fatigue effects were distributed. These revised versions of the tests are contained in Appendix H (limited distribution).

8. Construct Auxiliary Instruments

Task-specific Performance Rating Scales. The multi-method measurement approach applied to the 15 tasks which utilized two aspects of performance evaluation (hands-on and knowledge). To complete the approach, a qualitative evaluation was developed in which the soldier's direct or most immediate supervisors and the soldier's peers would be asked to rate his or her performance on those tasks measured also by hands-on and knowledge methods. These ratings, although the least direct of the measures, would permit the measurement of affective dimensions that could not feasibly be tapped by the other means. A 7-point rating scale was developed in which supervisors and peers were asked to rate each soldier tested on his or her performance of the 15 tasks, compared with other soldiers in the same MOS and skill level. Appendix I shows a sample of this rating form; the heading portion is shown in Figure 2.

During the Batch A field test of these scales, it was observed that supervisors and peers, given only the task title, may not have been sure of the tasks they were rating. Low interrater reliabilities supported this observation. Consequently, for the Batch B data collection, task statements were augmented with brief descriptions of the tasks, which had been developed for the task clustering phase of development (see Appendix D for a sample). This was done only for two MOS, the 31C Single Channel Radio Operator and the 19E Armor Crewman, in order to test the efficacy of the method.

Job History Questionnaire. Although soldiers in a given MOS share a common pool of potential tasks, their task experience may vary substantially. The most widespread reason for these differences is assignments latitude in the MOS. The assignments may be formal, as when an SL1 Armor Crewman is assigned as the driver or as the loader on the tank, or they may be informal divisions of labor, such as when one Administrative Specialist primarily types orders while another primarily types correspondence. A more extreme reason for task experience differences in an MOS occurs when soldiers are assigned to duties not typically associated with their MOS. For example, an Armor Crewman may be assigned to drive a 1/4-ton truck, or a Medical Specialist may perform clerical tasks. Such soldiers are not given Special or Additional Skill Identifiers, nor are they considered to be working in a secondary MOS; there is no code appended to their MOS that identifies them as working outside their primary MOS. They are simply tankers who drive trucks or medics who type and file. The likelihood of differences in task experience is further increased by differences in unit training emphasis where training schedules at battalion, company, and

RATING JOB TASK PERFORMANCE (64C)

Name _____ Date _____

Not all soldiers can perform all of their job tasks equally well. On this rating form we would like you to rate how well _____ can do each of the job tasks that follow.

Read each task statement carefully and think about how well this soldier can do the task: Among the very best you have seen at doing the task? Among the very worst? Somewhere in between? When you make up your mind, find the number on the scale below that describes your rating and circle that number after the task. If you can't rate a task because you haven't seen this soldier do it, circle a 0 for the task. But please do your best to rate the soldier on each task.

	1	2	3	4	5	6	7	0
	Among the Very Worst	Worse Than Most Others	Worse Than Many Others	About the Same As Other Soldiers	Better Than Many Others	Better Than Most Others	Among the Very Best	Cannot Rate
_____ Task _____								

Figure 2. Instructions to peers and supervisors for rating job task performance.

platoon level emphasize different tasks. As a result of these circumstances, soldiers' experiences vary, even within an MOS and location. This variance probably affects task performance.

Given that the central thrust of Project A is the validation of selection and classification measures, any differential task experience that affects performance is a contaminating variable. That is, if the differences in task experiences of sampled soldiers are wide enough to have an impact on task performance, experience effects may also be strong enough to mask predictor relationships with performance. In this case, measures of experience would need to be incorporated into validation analyses so that predictor-criterion relationships could be assessed independent of experience.

To be able to assess the probable impact of experience on task performance, and consequently on the Concurrent Validation strategies, a Job History Questionnaire was developed to be administered to each soldier. Specifically, soldiers were asked to indicate how recently and how frequently (in the preceding six months) they had performed each of the 30 tasks selected as performance criteria. A copy of the questionnaire for the 13B Cannon Crewman is included as Appendix J.

Chapter 3

CONDUCT OF FIELD TESTS

GENERAL PROCEDURE

Field testing was conducted in two phases--from March to September 1984 for the Batch A MOS and from February through April 1985 for the Batch B MOS. The field test locations and numbers tested in each location are shown in Table 3. In each MOS, incumbents were tested for two full days. The hands-on and knowledge task performance tests each required one-half day of participant time; the other two 4-hour blocks were used for administration of various rating scales (including the Task Performance Rating Scales), questionnaires (including the Job History Questionnaire), and measures of training achievement (these other measures are described in other Project A field test reports).

Table 3

Soldiers by MOS by Location for Field Testing

Location	MOS									
	11B	13B	19E	31C	63B	64C	71L	91A	95B	Total
Fort Hood							48		42	90
Fort Lewis	29		30	16	13			24		112
Fort Polk	30		31	26	26		60	30	42	245
Fort Riley	30		24	26	29		21	34	30	194
Fort Stewart	31		30	23	27			21		132
USAREUR	58	150	57	57	61	155		58		596
TOTAL	178	150	172	148	156	155	129	167	114	1369

At each site, an officer and two NCO from one of the supporting units were assigned to support the field test. The officer provided liaison between the data collection team and the tested units, and the NCO coordinated the acquisition of equipment and personnel. At each test site, a test

site manager from the project staff supervised all of the research activity and maintained the orderly flow of personnel through the data collection points.

Before any instruments were administered, each soldier was asked to read a Privacy Act Statement, DA Form 4368-R. Project staff then gave a brief introduction, stating the purpose of the project and emphasizing the confidentiality of the data. The soldiers were then asked to complete a Background Information Form.

After the introductory session, soldiers moved in groups of about 15 to either the hands-on testing, one of the knowledge test sessions, or a rating session. The order of administration of the measures was counterbalanced across groups and locations within MOS.

After soldiers reported for testing, their first- and second-line supervisors were identified (either by the soldiers themselves or as designated by the soldiers' units) and were notified of the scheduled supervisor rating session. Considerable flexibility was required in providing alternate sessions for supervisors, including offering evening and weekend times for individuals. Each session normally took two to three hours.

HANDS-ON TESTS

Field tests were conducted using unit NCO as scorers under the control of a Hands-On Manager (HOM) from the project. In addition to monitoring the overall hands-on test administration, the HOM performed the following specific duties:

- Scorer Training - For one or two days (depending upon the site) prior to test administration, the scorers received orientation and training. The training was specific to the set of tasks and involved actual practice administration of tests to other scorers acting as examinees. The HOM tailored the training to the scorer and the task. A copy of the scorer training materials is at Appendix K.

- Scorer Monitoring - During testing, the performance of the scorers was monitored to keep problems from arising and correct any difficulties that arose. Any problems with the tests were noted and, if possible, changed for subsequent administration. Test site variations and modifications were noted.

For the Batch A MOS, hands-on tests were administered in a test-retest mode to gather scorer and performance reliability data. This procedure was terminated for Batch B MOS; however, some shadow scoring (simultaneous scoring by the NCO scorer and a project test developer) was conducted in Batch B for the same purpose.

Depending on the task being measured, the location for testing was outside (vehicle maintenance, weapons cleaning) or inside (typing, administer injection). Scorers assigned to each test station ensured that

the required equipment was on hand and that the station was set up correctly, and then followed the procedures for administering and scoring the tests. As each soldier entered the test station, the scorer read the instructions aloud to the soldier and began the measure. The length of time a soldier was at the test station depended both on the complexity of the task and the individual's speed of performance.

KNOWLEDGE TESTS

Project staff members served as the test administrators for the knowledge tests. Times to complete each test booklet were recorded to assist developers in later reducing the 4-hour block for the field test to the 2-hour block for the Concurrent Validation. Sample test monitor instructions for Batch A and Batch B are shown in Appendix L.

The MOS-specific knowledge tests were grouped into four booklets of about seven or eight tasks per booklet, with each booklet requiring about 45 minutes to complete. The order in which the booklets were administered and the order of the tasks in each booklet were rotated. A break of 10 to 15 minutes was scheduled between booklets. As noted earlier, the purpose of dividing the material into separate booklets was to try to control the effects of fatigue and waning interest.

TASK PERFORMANCE RATING SCALES

The administration of the peer and supervisory rating scales, including the task performance scales, was preceded by a series of steps to identify peers and supervisors for each soldier. The ratings of task performance were designed around "rating units." Each rating unit consisted of the individual soldier to be evaluated, four identifiable peers, and two identifiable supervisors. A peer was defined as an individual soldier (from the group being tested) who had been in the unit for at least two months and had observed the ratee's job performance on several occasions. A supervisor was defined as the individual's first- or second-line supervisor (normally his rater and endorser.)

The procedure for assigning ratees to peer raters had two major steps:

1. A screening step in which it was determined which ratees could <u>potentially</u> be rated by which raters.

2. A computerized random assignment procedure which assigned raters to ratees within the constraints that (a) the rater indicated he/she could rate the ratee (Step 1); (b) ratees with few potential raters were given priority in the randomized assignment process; (c) the number of ratees assigned the various raters was equalized as much as possible across raters; and (d) the number of ratees any given rater was asked to rate did not exceed a preset maximum.

In each peer rating session, the potential raters were given an alphabetized list of the ratees (i.e., the full list of the soldiers being tested in the MOS at that location.) They were told the purpose of the ratings within the context of the research, the criteria (e.g., minimum length of period of working together) which they should use in deciding whom they could rate, and the maximum number of people they would be asked to rate. They were also told that assignments of ratees to raters would be accomplished randomly and that the randomization procedure would attempt to equalize, as much as possible, the number of ratees that any one rater would have to rate. The importance of their careful and comprehensive examination of the list of ratees in Step 1 was emphasized.

Because the use of different administrators at different sites could be expected to result in inconsistencies in administration, an important concern was that all raters face the same (or a very similar) rating task. The interaction between the rating unit and the administrator was a serious potential confounding factor. Lower or higher average ratings from some raters could be a result of different "sets" (i.e., "rate more severely" or "rate more leniently") provided by administrators handling the rating sessions, rather than a reflection of the true performance of the soldiers being rated.

A rater training program was conducted in an attempt to standardize the rating task during both peer and supervisor rating sessions. The rating scale administrator used a rater training guide to discuss the content of each effectiveness category, and pointed out common rating errors that should be avoided. The training content is described more fully in Pulakos and Borman (1985).

A second major thrust of the rater training program was to make it possible to obtain high quality ratings from both the peers and the supervisors without the raters having to do more than a minimum of reading. As much as possible, oral administration characterized the rating sessions.

Chapter 4

ANALYSES OF FIELD TEST DATA

The analyses performed had three major uses:

- To direct revisions to the hands-on and knowledge tests to improve the reliability of performance measurement.
- To assess the adequacy of the measures.
- To permit refinement of the measures for the Concurrent Validation.

ANALYSES TO IMPROVE RELIABILITY OF PERFORMANCE MEASUREMENT

Analyses directed at revising the measures to improve their reliability focused on knowledge test items and hands-on test step information as well as on results at the task level. Observations of the testing by test developers were essential in determining the revisions necessary to correct problems.

Knowledge Test Item Information

A multifaceted item analysis procedure was used to generate the information for evaluating test items. It operated independently on each task test, using knowledge test items as the unit of input. The output included:

- For each knowledge item, the number and percent who selected each alternative.
- For each item alternative, the Brogden-Clemans item-total correlation, where total score represents all items less the subject item in that task test.

The procedure also included options for multiple keying and zero-weighting of individual items, in anticipation of proposed revising or deleting of items.

Those items that were particularly easy (more than 95% pass) or particularly difficult (fewer than 10% pass), or that had low or negative item-total correlations were examined first for keying errors or obvious sources of cueing. Deficient items that could not be corrected were then deleted, and the item analysis was produced again. The process was iterative; various sets of items were analyzed, and the set that produced the highest coefficient alpha for the entire knowledge task test with an acceptable pass rate (between 15% and 90%) was retained.

However, exceptions to these criteria were made on a case-by-case basis. Items with extremely high or low difficulties provided relatively

few discriminations, yet some such items might be needed to enhance test acceptability because of the importance of the content tested or to preserve a single knowledge test of a Common Task tested across MOS. Items with low item-total correlations might be deficient in some respect or might simply be increasing test content heterogeneity. Since neither type of information provided conclusive evidence regarding an item's utility, both were applied in a judicious and cautious manner.

Hands-On Test Step Information

For each hands-on step, the number and percentage who scored GO and NO-GO were determined. The Brogden-Clemans biserial was computed for hands-on steps just as for knowledge test items.

Steps that had low or negative correlations with the test total score were reviewed to identify situations where performance could not be reliably observed or measured, or where performance that had been scored as NO-GO was in fact prescribed by local practices and was as correct at that site as doctrinally prescribed procedures. Instructions to scorers and to soldiers were revised as necessary to insure consistent scoring.

Use of step difficulty data to revise hands-on tests was limited by a number of considerations. First, a task test usually represents an integrated procedure; typically, each individual step must be performed by someone in order for the task to continue. Removal of a step from a score-sheet, regardless of its psychometric properties, might only confuse or frustrate the scorer. Second, removal of a step from a task test that has been developed on the basis of doctrine would often result in deleting a doctrinal requirement and undermining the credibility and acceptability of the hands-on measure.

Because of these considerations, very few performance measures were dropped from scoring on hands-on tests, regardless of their difficulty level. Under certain limited circumstances, exceptions were made. On a very few hands-on tasks, the test steps represent a sample of performance from a large domain (e.g., "Identify Threat Aircraft," "Recognize Armor Vehicles," "Give Visual Signals"); in such cases, individual steps could be deleted without damaging task coverage or test appearance. If discrete subtasks were, as a group, extremely easy or difficult, they could be dropped from the test if it would effect savings in time or equipment without sacrificing face validity. Very easy or difficult steps might be retained if they were scattered throughout the test, although easy steps were sometimes merged with subsequent steps. For example, in the task "Administer an Injection," the step "Removed protective cover from needle" was passed by all soldiers; the step was combined with the following step "Did not contaminate needle while removing cover," to read "Removed needle cover without contaminating needle."

Task Test Information

At the task test level, the field test results included task test means and standard deviations, and measures of test reliability.

Three types of reliability estimates were considered for use with hands-on task test data: test-retest, interscorer, and internal consistency. For reasons to be discussed below, only the internal consistency estimates were systematically used in test revision.

So that test-retest reliability could be computed, all soldiers in the Batch A MOS were retested on a subset of the same tasks they had been tested on initially. Due to scheduling and resource constraints, the interval between first and second testing was only two to seven days. Thus, memory of initial testing was a probable contaminant of retesting performance. Soldiers were aware that they would be retested and some were found to have trained to improve their performance between the two testing sessions; scores improved on second testing for many soldiers. Training during the interval was not consistent across soldiers, but varied partly as a function of motivation and partly as a function of the extent to which special duties may have restricted training opportunities. On the other hand, some soldiers resented having to repeat the test; some told the scorer that they were unfamiliar with the task, when in fact they had scored very high on initial testing. Thus, retest scores were contaminated widely and variably by motivational factors. Overall, test-retest data were of limited utility and were not collected for Batch B soldiers.

The use of alternate forms of a test offers an approximation of test-retest reliability. However, development and large-scale administration of alternate forms in either the hands-on or knowledge mode were beyond the resources of the project.

An attempt to acquire interscorer reliability estimates was made in Batch B by having a Project A staff member score the soldier at the same time the NCO was scoring. Two factors limited the feasibility of this approach. First, sufficient personnel were not available to monitor all eight stations within an MOS for the length of time necessary to accumulate sufficient data. The problem was exacerbated when, for whatever reason, two MOS had to be tested simultaneously. Second, for some MOS, particularly those performed in the radio-teletype rig for 31C, and in the tank for 19E, it was difficult or even impossible to have multiple scorers without interfering with either the examinee or the primary scorer. Because of these factors, interscorer reliability data were insufficient to systematically affect the process of revising task measures.

By process of elimination, the reliability measure of choice for the hands-on tests was an internal consistency estimate, using coefficient alpha. While internal consistency was, under the circumstances, the best measure available, it is far from ideal. First, if the content of a task is heterogeneous, the correlation will be low, regardless of the quality of the test. Second, an internal consistency estimate assumes independence of test items (here, task steps). For many hands-on tests, where final task steps cannot be completed if initial steps are missed, this condition cannot be met. Third, the measure is affected by test length, and hands-on task tests are short, generally ranging from 4 to 52 steps.

Task knowledge test reliability was measured in terms of coefficient alpha. As with hands-on tests, any internal consistency measure is affected by test heterogeneity, which is often an integral characteristic of the

task, and by test length, which varied on knowledge tests from 3 to 16 items. However, knowledge test items are psychometrically independent, unlike hands-on test steps.

Because of these considerations, internal consistency as an estimate of reliability was used cautiously, and in the context of other data regarding a task. A low correlation was a signal that the test deserved special attention; it was not considered, by itself, to be compelling evidence that the test was inadequate.

As an independent measure of performance each task test will indeed have a relatively large error component; thus, the most appropriate measure of soldier performance is the cumulated score across tasks.

ANALYSES TO ASSESS THE ADEQUACY OF THE MEASURES

Basic information was collected to assess each of the task-based and related measures administered in the field test. Type of information collected according to type of measure is summarized below.

Hands-On and Job Knowledge Tests

For each of these two kinds of tests, the results examined included mean percent score across tasks for each MOS, standard deviation of percent score, and split half reliability (using task test scores as the units for the split).

Task Performance Rating Scales

Inspection of the rating data revealed level differences in the mean ratings provided by two or more raters of the same soldier. Because interest centered on the profile for the individual soldier, the raters' responses were adjusted to eliminate these level differences. Additionally, a small number of raters were identified as outliers, in that their ratings were severely discrepant overall from the ratings of other raters on the same soldiers; their rating data were excluded from the analyses. (The procedures for adjusting the ratings for level effects and for identifying outliers are described in Pulakos & Borman, 1985.)

Means and standard deviations were computed on the adjusted ratings on each 7-point scale. Interrater reliabilities were estimated by means of intraclass correlations, and are reported as the reliability of two supervisors per soldier and four (Batch A) or three (Batch B) peer raters per soldier. The adjustment was made because numbers of raters varied for each soldier; it had seemed reasonable to expect that ratings could be obtained from two supervisors and four peers during the Concurrent Validation, but further experience in Batch B data collection suggested that three peers per soldier was more reasonable.

Intercorrelations Between Hands-On, Job Knowledge, and Task Performance Rating Scales

Correlations between hands-on and job knowledge tests were of particular interest because each type of measure had been designed as an integral component of an overall task-based measure. Correlations were calculated between these two types of measures for each MOS. Correlations were also computed across the MOS in four categories: clerical, skilled technical, operations, and combat. These categories of MOS have been identified by McLaughlin, Rossmeissl, Wise, Brandt and Wang (1984) on the basis of aptitudes measured by the Armed Services Vocational Aptitude Battery (ASVAB).

Correlations between task performance rating scales and the other task-based measures were also calculated, although these correlations were considered more exploratory than definitive.

Job History Questionnaire

Job history responses were analyzed to determine whether task experience as captured by the Job History Questionnaire is related to performance on the task-specific criterion measures. If a sufficient relationship were found, job history data would also be collected during the Concurrent Validation.

Because the Job History Questionnaire data analyses were performed solely to inform the decision on whether to continue collecting job history information, attention was focused on one Batch A MOS (13B) and three Batch B MOS (11B, 19E, and 63B). For 13B, task frequency and recency responses were summed to give a single index for each task, which was correlated with knowledge test scores and on hands-on test scores as an estimate of the overall effect of experience. For the Batch B MOS, simple correlations between the two kinds of job history responses and the two kinds of test scores were calculated. In all four MOS, job history means and standard deviations were also calculated, by task.

The 13B MOS presents an unusual case in that the decision had been made to test the crewmen of various types of howitzers, with weapon-specific versions of some task tests. Thus, two groups of 13B crewmen--M109 howitzer and M110 howitzer--were represented in the field tests. In order to analyze the results from these two groups as a single set of data, scores for the tracked tasks were standardized within their respective groups before the groups were merged.

ANALYSES TO REFINE THE MEASURES

In refining the set of hands-on and knowledge tests, the goal for each MOS was a reduction in knowledge test items of 25 - 40% (depending on the MOS), and a set of between 14 and 17 hands-on task tests. Experience during field testing indicated that the 2-hour knowledge test session and 4-hour hands-on test block allocated for Concurrent Validation could support such a

set. Each task test set was intended to provide the best task coverage with the strongest psychometric properties possible, whether in a hands-on or knowledge mode.

For these adjustments, the field test results used included judgments of hands-on test suitability, task test means and standard deviations, measures of test reliability (coefficient alpha or split-half reliability), and correlations between knowledge and hands-on tests.

Suitability Judgments

Although field test data can inform developers on issues of reliability, there remains the question of test validity. Hands-on tests of tasks often require compromises in large-scale testings: Conditions can be standardized, but realism may be sacrificed on some tasks; portions of tasks may not be tested because of equipment and safety constraints. To assess the effects of these compromises during the field tests, observers/developers rated the hands-on tests according to suitability for hands-on testing. The points to be considered were standardization of conditions, reliability of scoring, and quality of task coverage. Instructions for the review are at Appendix M.

The suitability judgments were used first in isolation to define the pool of hands-on tests. Tasks that were judged to be unsuitable (summed rating of 0, 1, or 2) by a majority of the observers were then dropped from hands-on testing. Hands-on tests that were field-tested in other MOS and judged as suitable were added to the set of available hands-on tests for each MOS where the task was tested in the knowledge mode (i.e., where the task was selected as one of the 30 tasks).

Task Test Information

The suitability ratings were then used with other data to further refine the task test sets as needed. First, if the hands-on test set was too long (more than 17 tests, or likely to run over 3 hours) after revisions, developers dropped hands-on tests that were not very suitable for the hands-on mode (summed ratings of 3 or 4), or that were suitable but had high correlations (over .40) with strong knowledge counterparts, or that overlapped with similar skilled psychomotor hands-on tests. However, if dropping such tests would not have effected a savings, because the tests were not time-consuming or resource-intensive, they were often retained. When the hands-on set comprised 14-17 of the best available hands-on tests, the set was considered final.

If, after revisions, the knowledge test set had 60 - 75% as many items as before, the tests were considered feasible for the 2-hour time slot. The knowledge test set was then accepted as complete, and finalized for Proponent review.

However, if there were still too many items, the strengths and weaknesses of each hands-on and knowledge test were examined in an attempt to determine which test mode was best measuring each element of each task. The

rationale behind this procedure stems from two lines of guidance: first, that the set of knowledge items needed to be reduced, and second, that as far as possible every task should be represented in the knowledge tests. After detecting items that were not reliably measuring the soldiers' knowledge of the task, further reductions focussed on ensuring that the items which tested cognitive components be retained in the knowledge mode, while the hands-on mode would, where possible, cover the skilled components of the task. If the knowledge test was not too long, of course, task elements could be covered in both modes of testing. The steps in the procedure are detailed in Appendix N.

The procedure considers whether or not the test was or was not revised after the field test, test difficulty, variability in scores, reliability, and hands-on suitability. This knowledge test information was considered in conjunction with an analysis of the specific content overlap between hands-on and knowledge tests and with the statistical correlation between hands-on and knowledge tests. Knowledge tests were gradually reduced to items that were demonstrated to be reliable measures of the tasks, by considering first the items needing revision or with lower statistics, and then the stronger knowledge test items that were found to be redundant (by overlap or correlation) with hands-on tests. The process was carried out one task at a time, until the number of knowledge items remaining was reduced sufficiently for the Concurrent Validation.

Chapter 5

RESULTS OF FIELD TESTING

IMPROVING RELIABILITY OF PERFORMANCE MEASUREMENT

Knowledge Test Revisions

Revisions were made on between 14 and 18% of the items in each MOS set; between 17 and 24% of the items were dropped. For the most part, the revisions were made on Common Task tests that had been developed for Batch A MOS, and that had been selected for testing in one or more Batch B MOS so the effect of the revisions could be assessed.

Hands-On Test Revisions

Very few performance measures were dropped from scoring on hands-on tests. Almost every test had at least minor wording changes in instructions or steps.

ASSESSING THE ADEQUACY OF THE MEASURES

Knowledge Task Tests

Following deletion of weak items (and holding out of analysis items slated for revision), the distributions of items with regard to difficulty and item-total correlations for each of the nine MOS were as summarized in Table 4. (Distributions are displayed in Appendix O.) For all MOS but three, the difficulty level (percent passing) mode was in the 41 to 60% bracket; for the 91A (Medical Specialist), 19E (Armor Crewman), and 95B (Military Police), the mode was between 81% and 100%. The median difficulty levels were 55% to 58% for five of the MOS, with the 63B (Light Wheel Vehicle Mechanic) as well as 91A, 19E, and 95B tests having medians of 65% to 74%. Although some skew in item difficulties was observed, it was not extreme.

The item-total correlation distributions were also highly similar across the nine MOS, with most items exhibiting correlations of .21 to .40 in each MOS. Pruning items on the basis of low correlations was done very conservatively, especially in cases where items behaved well in most of the MOS where the tasks were tested. As a result, there remained in each knowledge component items with low or negative correlations with the task total score; these ranged from 9% of the items in the 13B (Cannon Crewman) tests to 29% in the 19E (Armor Crewman) tests with correlations below .20. Negative correlations were found in no more than 8.8% of the items in any of the nine MOS. The average of the item-total correlations in the various knowledge components ranged from .30 to .38.

Table 4

Summary of Item Difficulties (Percent Passing) and Item-Total Correlations for Knowledge Components in Nine MOS

MOS	Number of Items		Mean	Median	Min	Max
13B Cannon Crewman	236	Difficulty(%)	59.2	55.5	13.4	97.2
		Item-Total(r)	.38	.38	-.06	.88
64C Motor Transport Operator	166	Difficulty(%)	60.7	58.0	03.6	94.3
		Item-Total(r)	.31	.32	-.00	.91
71L Administrative Specialist	170	Difficulty(%)	57.4	56.5	04.7	96.1
		Item-Total(r)	.30	.31	-.19	.84
95B Military Police	177	Difficulty(%)	66.4	74.0	00.0	100.0
		Item-Total(r)	.33	.32	.00	.82
11B Infantryman	228	Difficulty(%)	57.3	55.4	05.3	97.1
		Item-Total(r)	.30	.31	-.39	.88
19E Armor Crewman	205	Difficulty(%)	64.6	66.8	13.4	96.9
		Item-Total(r)	.32	.31	-.26	.95
31C Single Channel Radio Operator	211	Difficulty(%)	58.0	57.1	11.3	95.4
		Item-Total(r)	.31	.31	-.09	.84
63B Light Wheel Vehicle Mechanic	197	Difficulty(%)	65.1	64.5	07.8	97.4
		Item-Total(r)	.30	.30	-.13	.92
91A Medical Specialist	236	Difficulty(%)	66.9	69.0	08.6	98.7
		Item-Total(r)	.32	.32	-.25	.78

The means, standard deviations, and reliabilities across the tests in each MOS are shown in Table 5 (individual test means are presented in Appendix P); the reliabilities are split-half coefficients, using 15 tests in each half, corrected to a total length of 30 tests.

For all MOS, the majority of task means were between about 35% and 85%; overall knowledge component means (the mean of the task means) were from 55 to 70%. The standard deviations were also similar across the nine MOS, and although coefficient alphas were variable across tasks, split-half reliabilities were in the .70s to .90s for the full knowledge components.

Table 5

Means, Standard Deviations, and
Split-Half Reliabilities for
Knowledge Test Components for Nine MOS

MOS	Mean	Standard Deviation	Split-Half Reliability[a]
13B - Cannon Crewman	58.9	12.6	.86
64C - Motor Transport Operator	60.3	10.1	.79
71L - Administrative Specialist	55.8	10.4	.81
95B - Military Police	66.4	9.2	.75
11B - Infantryman	56.0	10.5	.91
19E - Armor Crewman	64.0	10.1	.90
31C - Single Channel Radio Operator	57.7	9.6	.84
63B - Light Wheel Vehicle Mechanic	64.4	9.1	.86
91A - Medical Specialist	69.8	8.1	.85

[a]Fifteen task tests in each half, corrected to a total length of 30 tests.

The reliabilities (coefficient alpha) of tests appearing in multiple MOS are shown in Table 6. These reliabilities are reasonably consistent across MOS, though with occasional outliers. The actual magnitude of the correlations is, for many individual tests, disappointing. However, some of the tests are very short, no more than 3-5 items.

When knowledge tests are combined into knowledge components, the reliabilities generally increase over those of individual tests. For a number of analyses, however, the individual tests will be expected to stand alone, so a careful review of the reliabilities of the measures will be in order when data analysis of the Concurrent Validation is undertaken.

Hands-On Task Tests

Table 7 shows, for each MOS, the mean, standard deviation, and reliability estimate (coefficient alpha) of the hands-on component across revised task tests (statistics for each hands-on test are presented in Appendix Q). The mean scores on tasks in each MOS mostly fall between 40% and 80%, although each MOS has a few tasks that are very difficult and a few

Table 6

Coefficient Alpha of Knowledge Tests
Appearing in Multiple MOS

Test	13B	64C	71L	95B	11B	19E	31C	63B	91A
Perform CPR	31	34		38	33	38	41		55
Administer nerve antidote to self	55		39		36				
Prevent shock	22					12			31
Put on field dressing		34	39	39	19	15	31	16	31
Administer nerve agent antidote to buddy		58						32	
Load, reduce stoppage, clear M16	56	46	47	52			51	32	43
Perform operator maintenance on M16		31	38		39		44	22	
Load, reduce, clear M60		30		40	47				
Perform operator maintenance .45				45		36			
Determine azimuth using a compass			81	84				74	
Determine grid coordinates		23	53	57		79	74	70	74
Decontaminate skin	71	42		48			47		47
Put on M17 mask	50	49	44	56	49			33	
Put on protective clothing	56	55	31		40	31	52	39	40
Maintain M17 mask			38	53			28		
Challenge and Password	46	48						41	
Know rights as POW			48			45	44		
Noise, light, litter discipline			38				12		07
Move under fire				59	56				
Identify armored vehicles	62			64	68	75	57		58
Camouflage equipment	31	31						08	
Camouflage self			06	47	48				
Report information - SALUTE		76			84	74			
Operate vehicle in convoy		40		36					

Table 7

Means, Standard Deviations, and Split-Half Reliabilities for Hands-On Components for Nine MOS

MOS	N	Mean %	Standard Deviation	Split-Half Reliability
13B - Cannon Crewman	146	54.5	14.0	.82
64C - Motor Transport Operator	149	72.9	9.1	.59
71L - Administrative Specialist	126	62.1	9.9	.66
95B - Military Police	113	70.8	5.8	.30
11B - Infantryman	162	56.1	12.3	.49
19E - Armor Crewman	106	81.1	11.8	.56
31C - Single Channel Radio Operator	140	80.1	10.7	.44
63B - Light Wheel Vehicle Mechanic	126	79.8	8.7	.49
91A - Medical Specialist	159	83.4	11.4	.35

[a]Calculated as 8-test score correlated with 7-test score, corrected to 15 tests.

tasks that are very easy for most of the soldiers tested. The standard deviations for task tests are in many cases high relative to the means. This is at least in part an artifact of the sequential nature of many of the hands-on tests: If soldiers cannot perform early steps, the test stops and remaining steps are failed.

The reliabilities shown for the individual task tests range from about .40 to .90 for most tasks. The very high figures tend to occur on tasks where the performance measures define a sequence of behaviors; on these tasks, soldiers perform all steps up to some point and none after that point, resulting in reliability estimates that are spuriously high. For most MOS, the overall split halves, calculated using seven scores against eight scores (odd-even, using the orders shown in Appendix Q), are noticeably lower, but these may be underestimates, as the two forms are arranged from tests of heterogeneous tasks.

A number of the hands-on tests appeared in Batch B MOS as well as in Batch A. As previously discussed, internal consistency estimates of hands-on test reliability are inflated in an absolute sense, but are useful for purposes of comparison. Table 8 shows that the reliability estimates

Table 8

Reliability (Coefficient Alpha) of Hands-on Tests Appearing in Multiple MOS

Test	Reliability		
Perform CPR	.94 (95B)	.91 (91A)	
Put on field/pressure dressing	.76 (71L)	.68 (91A)	
Load, reduce, clear M16	.61 (95B)	.65 (31C)	
Load, reduce, clear M60	.98 (64C)	.95 (11B)	
Perform operator maintenance on M16	.82 (71L)	.92 (11B)	
Put on protective clothing	.90 (64C)	.88 (31C)	
Perform PMCS	.56 (64C)	.66 (95B)	.74 (31C)

for these tests were fairly consistent across MOS. The task that exhibits the least consistency, "Perform PMCS (Preventive Maintenance Checks and Services)," was performed on different vehicles in the three MOS and thus some scoring points differed across the three MOS involved.

Task Performance Rating Scales

Summary statistics on the task performance rating scales across the 15 tasks in each MOS are presented in Table 9; results by task are in Appendix R. The distributions for the rating scales were surprisingly free of leniency and skewness, with task means mostly between 4 and 5 on the 7-point scale and standard deviations mostly between .80 and 1.10.

Reliabilities varied widely across the tasks. In MOS such as 71L (Administrative Specialist), where soldiers work in isolation from each other or with only one or two others, few peer ratings were obtained on each soldier and reliabilities are correspondingly lower. On the other hand, among 11B Infantrymen, the mean number of peer ratings was higher; many of the soldiers comprised training cohorts, who had been together since their earliest Army training. Some tasks that soldiers rarely perform were also characterized by lower numbers of ratings and lower reliabilities.

Intercorrelations Among Task-Based Measures

For each of the nine MOS, performance on 15 tasks was assessed by four methods: hands-on tests, knowledge tests, supervisor ratings, and peer

Table 9

Means, Standard Deviations, Number of Raters, and Interrater Reliabilities of Supervisor and Peer Ratings Across 15 Tasks for Nine MOS

MOS	Group	Mean Raters	Mean[a]	Standard Deviation[a]	Interrater Reliability[b]
13B - Cannon Crewman	Sup.	1.5	4.99	.72	.67
	Peer	2.5	4.85	.60	.87
64C - Motor Transport Operator	Sup.	1.8	4.35	.64	.69
	Peer	2.6	4.26	.58	.70
71L - Administrative Specialist	Sup.	1.0	4.97	.70	.75
	Peer	1.9	4.97	.64	.60
95B - Military Police	Sup.	1.9	4.51	.49	.64
	Peer	3.4	4.53	.46	.82
11B - Infantryman	Sup.	1.8	4.45	.59	.74
	Peer	3.0	4.50	.55	.77
19E - Armor Crewman	Sup.	1.7	4.69	.62	.76
	Peer	3.0	4.71	.45	.67
31C - Single Channel Radio Operator	Sup.	1.7	4.68	.68	.81
	Peer	2.5	4.68	.58	.74
63B - Light Wheel Vehicle Mechanic	Sup.	1.8	4.72	.68	.76
	Peer	2.1	4.68	.63	.81
91A - Medical Specialist	Sup.	1.6	4.97	.75	.69
	Peer	3.1	4.95	.60	.81

[a]Computed on adjusted ratings.

[b]Computed on adjusted ratings; corrected to reliabilities for two supervisors and four (Batch A) or three (Batch B) peers.

ratings. Thus, a 60x60 correlation matrix could be generated for each MOS, as a multimethod-multitrait matrix (where traits are tasks). For purposes of simple examination each MOS matrix was collapsed, by averaging correlations across tasks, to a 4x4 method matrix (see Figure 3). For each pair of methods, the 15 correlations between the two methods on the same tasks (heteromethod-monotrait) were averaged and are shown above the diagonals of the method matrixes. The 210 correlations between each pair of methods on

13B - Cannon Crewman

	HO	K	R-S	R-P
HO	16	20	18	23
K	10	21	15	16
R-S	10	12	59	31
R-P	16	10	24	53

11B - Infantryman

	HO	K	R-S	R-P
HO	12	24	18	16
K	11	19	15	12
R-S	11	10	28	35
R-P	10	09	24	31

64C - Motor Transport Operator

	HO	K	R-S	R-P
HO	07	14	11	08
K	07	09	12	05
R-S	07	05	48	38
R-P	05	02	30	35

19E - Armor Crewman

	HO	K	R-S	R-P
HO	13	14	09	09
K	10	19	10	04
R-S	07	04	30	28
R-P	03	04	18	27

71L - Administrative Specialist

	HO	K	R-S	R-P
HO	11	18	12	05
K	10	15	10	06
R-S	05	03	36	44
R-P	04	02	35	46

31C - Single Channel Radio Operator

	HO	K	R-S	R-P
HO	23	21	14	15
K	12	14	09	13
R-S	05	06	41	41
R-P	10	08	32	43

95B - Military Police

	HO	K	R-S	R-P
HO	03	12	11	10
K	01	09	06	06
R-S	05	04	37	35
R-P	05	03	26	39

63B - Light Wheel Vehicle Mechanic

	HO	K	R-S	R-P
HO	18	13	06	08
K	07	13	04	11
R-S	05	03	46	36
R-P	03	06	26	36

LEGEND:

	Hands-On	Knowledge	Ratings Supervisor	Peer
Hands-On			Different method, same task	
Knowledge				
Ratings-Supervisor			Same method, different task	
Ratings-Peer	Different method, different task			

91A - Medical Specialist

	HO	K	R-S	R-P
HO	11	11	08	12
K	05	15	04	00
R-S	03	-01	45	37
R-P	05	-01	27	43

Figure 3. Average correlations between task measurement methods on same tasks and different tasks for nine MOS.

different tasks (heteromethod-heterotrait) were averaged and entered below the diagonals of the method matrixes. Finally, the 105 correlations between pairs of tasks measured by the same method (monomethod-heterotrait) were averaged and are shown in the diagonals of the method matrixes.

In general, there are three considerations in examining a full multimethod-multitrait matrix: (1) The heteromethod-monotrait correlations (above the diagonals) are indications of convergent validity among the methods, the extent to which the different methods measure the same trait (here, the traits are proficiency on tasks). (2) These same validity coefficients (above the diagonals) should be greater than the corresponding values in the heteromethod-heterotrait triangle (below the diagonals), as an indication that the method-trait relationships are not all a reflection of some other unspecified factor. (3) The monomethod-heterotrait correlations (in the diagonals) should be lower than the coefficients above the diagonal, as evidence of discriminant validity--that is, the methods of measuring tasks are not overshadowing differences among tasks.

Without exception, the average correlations are highest both between and within peer and supervisor ratings, with method variance (different tasks) in general higher than variance accounted for by tasks. For hands-on and knowledge tests, the average of same-task correlations between the two methods (above the diagonal) was higher than either of the single-method different-task average correlations (in the diagonal), which were in turn usually higher than the average correlation between the two methods on different tasks (below the diagonal). The lower correlations between the task tests and task ratings, even on the same tasks (above the diagonal), further evidences the preponderant influence of the rating method.

Just as reliabilities are higher for each measurement method across tasks, the correlations among the methods tend to be higher when results are aggregated across tasks to the component level (see Figure 4). Again, the correlations between the two rating methods are highest, and correlations between rating methods and test methods are in general lowest. The exceptions are among 95B (Military Police) where the hands-on/knowledge correlation was particularly low, and among 11B (Infantryman) where ratings and test results were correlated nearly as highly as the two test methods.

Table 10 shows overall correlations between hands-on and knowledge tests for MOS grouped by occupational category. The categories used correspond to Aptitude Area composites identified by McLaughlin, et al. (1984), based on which ASVAB tests were most predictive of future training performance success for particular Army MOS. These composites were labeled: clerical, operations, combat, and skilled technical. The correlations were clearly lowest in the skilled technical category; otherwise, there were no major differences between groupings.

Job History Questionnaire

In analyzing the job experience statistics for 13B, Cannon Crewman (the Batch A MOS administered this questionnaire), recency and frequency were summed with frequency reverse scored prior to summing. Thus, high scores

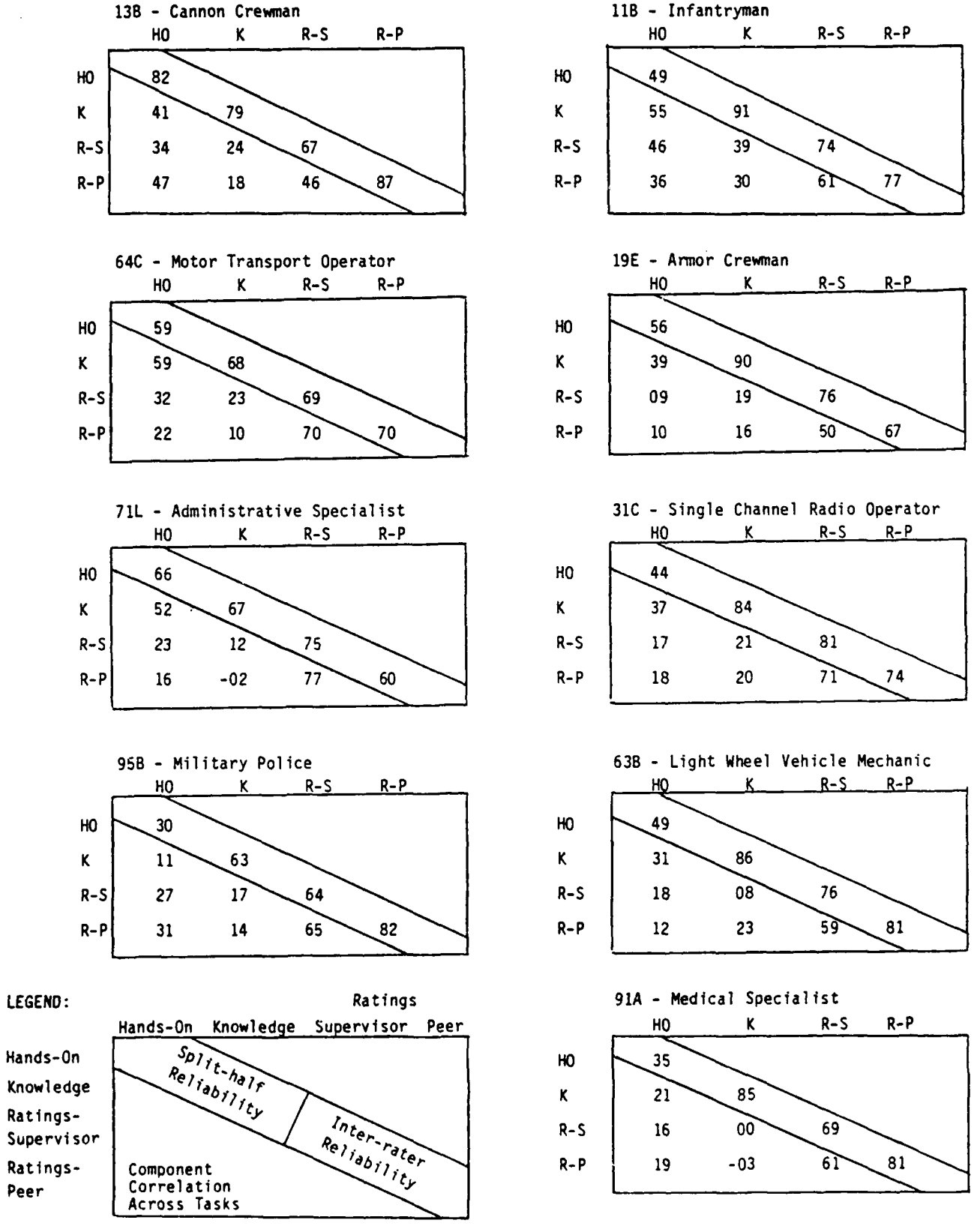

Figure 4. Reliabilities and correlations between task measurement methods across tasks for nine MOS.

Table 10

Correlations Between Hands-On and Knowledge Test Components for MOS Classified by Type of Occupation

Type of Occupation (MOS)	Total Sample Size	Correlation Between Knowledge and Hands-On	
		r^a	Corrected r^b
Clerical (71L-Administrative Specialist)	126	.52	.76
Operations (63B-Light Wheel Vehicle Mechanic; 64C-Motor Transport Operator; 31C-Single Channel Radio Operator)	393	.43	.71
Combat (11B-Infantryman; 13B-Cannon Crewman; 19E-Armor Crewman)	414	.46	.67
Skilled Technical (95B-Military Police; 91A-Medical Specialist)	250	.17	.35
OVERALL	1183	.39	.62

[a]Correlation between knowledge and hands-on test scores averaged across samples.

[b]Correlation between knowledge and hands-on test scores averaged across samples and corrected for attenuation.

indicate greater recency and/or frequency of task experience (see Appendix S, Table S.1). This summated experience score was significantly related, in the positive direction, with test scores for 9 of the 15 hands-on tests, and for 9 of the 30 knowledge tests. For six tasks, experience was significantly related to both knowledge and hands-on test performance. Results for this Batch A MOS certainly support the continued examination of job experience effects.

For the three Batch B MOS administered this questionnaire, frequency and recency were treated separately. Appendix Table S.2 presents the correlations between the job experience indexes and test performance for MOS 11B, Infantryman. Recency or frequency or both correlate significantly, and in the appropriate direction, for 7 of the 15 hands-on tests, and for 15 of the 32 knowledge tests. For six tasks, one or both experience indexes were related to both hands-on and knowledge performance.

Appendix Table S.3 presents statistics for MOS 19E, Armor Crewman, and Table S.4 presents statistics for MOS 63B, Light Wheel Vehicle Mechanic. For 19E, experience indexes were related to only three hands-on tests and to two knowledge tests; for one task, experience was significantly related to both knowledge and hands-on scores. For 63B, experience was significantly related to only two hands-on tests and to five knowledge tests, with none of the tasks having significant relationships with experience measures for both types of tests. For soldiers in these two MOS, experience differences appear to have less influence on performance.

REFINING TASK MEASURES

After initial revisions were made to the hands-on and knowledge tests to improve the reliability of performance measurement, the field test data and direct observations of testing were used to adjust the set of task measures to permit testing within the constraints of the Concurrent Validation resources.

The extent of the changes made on the tests, considering both obtained data and informed judgments, was small. Among Common Task tests, judgments of hands-on suitability resulted in deleting seven tests ("Recognize Armored Vehicles," "Visually Identify Threat Aircraft," "Decontaminate Skin," "Move Under Direct Fire," "Collect and Report Information," "Navigate on the Ground," and "Estimate Range"). Additionally, for each MOS one four MOS-specific tasks were dropped as not suitable for hands-on testing.

For each MOS, the set of hands-on tests, including those field-tested in other MOS and tests later judged not suitable, comprised 19 to 23 tasks; after suitability cuts were made, the hands-on sets were reduced to 15 to 19 tasks in each MOS. Appendix T lists the full set of hands-on tests that were developed and field tested for all MOS, and indicates which tests were dropped subsequently as unsuitable. For Common Tasks, the Appendix also indicates the other MOS for which the tasks were selected and where, therefore, they might also be tested hands-on.

After the weak items had been removed from the knowledge tests, all MOS except four (95B-Military Police, 31C-Single Channel Radio Operator, 63B-Light Wheel Vehicle Mechanic, 91A-Medical Specialist) required further knowledge test reduction. The procedure described earlier was followed to release the knowledge items that covered task elements better tested in the hands-on mode. From five to ten task tests per MOS were reduced in this fashion; some of those reductions resulted in tasks being tested only in the hands-on mode.

Table 11 summarizes the various adjustments to hands-on and knowledge tests for each of the nine MOS, in preparation for the Proponent agency review. A list of the tests to be reviewed for each MOS is presented in Appendix U.

Table 11

Summary of Adjustments[a]
to Hands-On and Knowledge Tests
Before Proponent Review

MOS	Hands-On Mode[b] Tests Dropped (A)	Tests Added (B)	Knowledge Mode[c] Items (C)	Items Dropped (D)	Items to HO (E)	Tests to HO (F)	Avg % to HO (G)	Tasks To Be Tested HO Only	HO & K	K Only
13B-Cannon Crewman[d]	3	5	279, 278	57, 55	45, 42	9	63% 62%	2	17	13
64C-Motor Transport Operator	5	6	265	66	31	8	71%	2	14	14
71L-Administrative Specialist	1	1	235	34	53	10	74%	5	10	15
95B-Military Police	4	4	281	71	0	-	-	0	15	15
11B-Infantryman	3	3	272	51	23	5	73%	2[e]	13	16
19E-Armor Crewman	2	2	252	40	16	5	46%	1	14	15
31C-Single Channel Radio Operator	2	2	245	30	0	-	-	0	15	15
63B-Light Wheel Vehicle Mechanic	0	0	248	52	0	-	-	0	15	15
91A-Medical Specialist	1	1	284	50	0	-	-	0	15	15

[a]Adjustments indicated by column headings are:
(A) Hands-on task tests dropped as not suitable.
(B) Hands-on task tests field-tested in other MOS.
(C) Number of knowledge items field-tested.
(D) Number of knowledge items dropped as unreliable.
(E) Items deleted from knowledge tests as better covered in hands-on mode.
(F) Number of tests from which items were deleted as better covered in hands-on mode.
(G) Average percent of items per test that were deleted as better covered in hands-on mode (for the tests in column (F)).

[b]Every MOS had 15 task tests in hands-on set for field test.

[c]Every MOS had 30 task tests in knowledge set for field test.

[d]Two versions of the knowledge test were prepared, for M109 and M110 howitzer crewmen.

[e]One task was developed for hands-on testing only.

Chapter 6

PROPONENT AGENCY REVIEW

The final step in the development of hands-on and knowledge tests was Proponent agency review. This step was consistent with the philosophy of obtaining input from Army subject matter experts at each major developmental stage and also was considered important with respect to the credibility of the measures developed.

A letter from the Deputy Chief of Staff for Training, Training and Doctrine Command, was sent to the Commanding Office at each Proponent agency, asking for a review of the performance measures that had been developed. Subsequently, project staff briefed Proponent representatives on the purpose and background of the project and the development of the performance measures. The Proponent was asked to consider two questions: (1) Do the measures reflect doctrine accurately, and (2) do the measures cover the major aspects of the job? A Proponent representative was given copies of the measures; staffing of the review was left to the discretion of the agency.

In general, considerable deference was given to the Proponent judgments; however, certain potential conditions were identified where strict adherence to such judgments could be counterproductive. One such conditions would be if Proponent changes, either to the content of large numbers of items within tasks or to the task list itself, were so extensive that the content of the measures were substantially altered. In practice, Proponent item changes generally occurred in much fewer than 10% of the items within an MOS, and most such changes involved the wording, not the basic content, of the item.

Changes affecting the task list occurred in only three MOS. Proponent comments and resulting actions may be summarized for each of these MOS as follows:

11B - Infantryman. The Infantry Center indicated that the primary emphasis for infantry should be nonmechanized. To that end, they advised dropping three tasks: "Perform PMCS on Tracked or Wheeled Vehicle," "Drive Tracked or Wheeled Vehicle," and "Operate as a Station in a Radio Net." Two tasks field-tested in other MOS were substituted: "Move Over, Through, or Around Obstacles," and "Identify Terrain Features on a Map." The Center also concurred with the addition of a hands-on test of the task, "Conduct Surveillance Without Electronic Devices"; the hands-on test of "Estimate Range" had been dropped in exchange. The 11B test set then included 30 tasks, 14 tested hands-on.

71L - Administrative Specialist. The Soldier Support Center, proponent for 71L, recommended that "Post Regulations and Directives" and "Perform CPR" be eliminated from the 71L task list. They also recommended that four tasks originally designated for testing in the knowledge mode be tested in the hands-on mode as well: "File Documents/Correspondence," "Type a Joint Message Form," "Type a Military Letter," and "Receipt and Transfer Classified Documents." To allow testing time for the additions, three

tasks, originally to be tested in both the hands-on and knowledge modes, will now be tested only in the knowledge mode: "Put On/Wear Protective Clothing (MOPP)," "Load, Reduce Stoppage and Clear M16A1 Rifle," and "Determine Azimuth with Lensatic Compass." These changes resulted in a 71L test set composed of 28 tasks, 14 tested in a hands-on mode.

95B - Military Police. The Military Police School, Proponent for 95B, indicated that the role of the military police was shifting toward a more combat-ready, rear area security requirement, rather than the domestic police role emphasized by the tasks selected for testing. They recommended that five tasks be added. Three of these, "Navigate from One Position on the Ground to Another Position," "Call for and Adjust Indirect Fire," and "Estimate Range," had previously been field-tested with 11B soldiers. Both hands-on and knowledge tests for these tasks were added. Another, "Use Automated CEOI," had been field-tested with 19E soldiers; this task was added to the list of knowledge tests only. The fifth task, "Load, Reduce a Stoppage, and Clear a Squad Automatic Weapon," not previously field-tested, was also added to the list of knowledge tests only. Four tasks were dropped. Two, "Perform a Wall Search" and "Apply Hand Irons," had initially been proposed for both hands-on and knowledge testing. The remaining two, "Operate a Vehicle in a Convoy" and "Establish/Operate Roadblock/ Checkpoint," had been on the knowledge only task list. The modified 95B test set consisted of 31 tasks, 16 tested in a hands-on mode.

In determining whether any of these task list changes constituted a major shift in content coverage, special consideration was given to the principle applied in the initial task selection process that every cluster of tasks be represented by at least one task. What impact did the Proponent changes have with respect to this principle? For 71L and 95B, each cluster was still represented after the Proponent changes had been implemented. For 11B, the deletion of "Perform PMCS on Tracked or Wheeled Vehicle" and "Drive Tracked or Wheeled Vehicle" left one cluster, consisting of tasks associated with vehicle operation and maintenance, unrepresented. However, since it was the Infantry School's position that tasks in this cluster did not represent the future orientation of the 11B MOS, this omission was considered acceptable.

A second condition where strict adherence to Proponent suggestions was not necessarily advisable was where the suggestions could not be easily reconciled with documented Army doctrine. Where conflict with documentation emerged, the discrepancy was pointed out; if the conflict was not resolved, items were deleted. Finally, if Proponent comments seemed to indicate a misunderstanding of the purpose or content of the test items, clarification was attempted. The basic approach was to continue discussions until some mutually agreeable solution could be found.

The Army Research Institute, with the concurrence of its General Officer Advisory Group, took the position that the performance measures would not be used to assess the validity of the predictor measures until the Proponent agencies provided concurrence with respect to the doctrinal accuracy and job coverage of such performance measures. For seven of the MOS, 11B (Infantryman), 13B (Cannon Crewman), 63B (Light Wheel Vehicle Mechanic), 64C (Motor Transport Operator), 71L (Administrative Specialist), 91A (Medical Specialist), and 95B (Military Police), Proponent agencies have

provided such concurrence by signed letter. For the remaining MOS, 19E (Armor Crewman) and 31C (Single Channel Radio Operator), changes have been made in response to Proponent comments and telephonic assurance has been obtained from a Proponent representative that no additional revisions beyond those initially specified will be requested. For these MOS, steps to obtain formal written concurrence from the Proponent are in progress.[1]

Copies of all tests, reflecting revisions based on field test data adjustments to fit constraints of Concurrent Validation, and changes recommended by Proponent agencies, are presented as Appendix V (limited distribution). The final array of tasks by test mode for each MOS is shown in Table 12.

Table 12

Final Array of Tasks Per Testing Mode for Concurrent Validation

MOS	Total Tasks	Hands-On Only	Hands-On And Knowledge	Knowledge Only	Knowledge Items
13B - Cannon Crewman	30	0	17	13	178,181
64C - Motor Transport Operator	30	2	14	14	167
71L - Administrative Specialist	28	3	11	14	144
95B - Military Police	31	0	16	15	213
11B - Infantryman	30	2	12	16	197
19E - Armor Crewman	30	1	14	15	192
31C - Single Channel Radio Operator	30	0	15	15	206
63B - Light Wheel Vehicle Mechanic	30	0	15	15	194
91A - Medical Specialist	30	0	15	15	229

[1]Status as of 31 December 1985.

Chapter 7

DISCUSSION

The results of the development effort, from the first perusals of the MOS task domains, through task selection, test development, and field test data collection, to the final production of criterion measures for Concurrent Validation, are impressive and satisfying at several levels. More than 200 knowledge tests and more than 100 hands-on tests were developed and field tested, and the field test experience was applied to the production of criterion measures of more than 200 tasks for the nine MOS. The tests provide broad coverage of each MOS in a manner which is both psychometrically sound and appealing to MOS Proponents.

Initial predictions of the capability of Army units to support hands-on tests and the ability of SL1 soldiers to comprehend the knowledge tests and rating scales were largely borne out during data collection. Where any serious misjudgments had been made in preparing materials, it was possible to effect corrections to eliminate the problems encountered.

The several methodologies developed for defining the task domains, obtaining SME judgments of task criticality and difficulty, selecting tasks for testing, assigning tasks to test modes, and reducing test sets to manageable arrays proved both comprehensive and flexible. The peculiarities of each MOS required that the methods be adapted at various points, yet for every MOS all vagaries were dealt with to the satisfaction of both developers and Proponents.

In general, means and standard deviations revealed a reasonable level of performance variability on hands-on and knowledge tests. In one MOS where the variability of hands-on tests was most limited, the 95B Military Police, there have been Proponent-directed changes which may result in increased variability in Concurrent Validation testing.

It would not be appropriate to interpret the means of any of the measures as an indicator of soldier quality. These were draft versions of tests which were administered for the purpose of determining what test revisions might be needed. For this purpose, no standard of acceptable performance had been identified. In the absence of such a standard, no conclusion about the quality of our enlisted soldiers based on these test scores would be meaningful.

To know whether correlations in this effort are high or low, some frame of reference is needed. Rumsey, Osborn, and Ford (1985) reviewed 19 comparisons between hands-on and job knowledge tests. For 13 of the 19 comparisons using work samples classified as "motor" because the majority of tasks involved physical manipulation of things (see Asher & Sciarrino, 1974, for a distinction between "motor" and "verbal" work samples), a mean correlation was found of .42 prior to correction for attenuation and .54 following such correction. Results were further divided into occupational categories, based primarily on which aptitude areas on the ASVAB, a set of cognitive tests, best predicted performance for that category; the categories were skilled technical, operations, combat arms, and electronics.

Table 13 shows corrected and uncorrected correlations in each of these categories. An additional category, clerical, was identified, but no investigations using a motor work sample had reported any results in this category.

Table 13

Reported Correlations Between
Hands-On (Motor) and Knowledge Tests

	Correlation	
	r[a]	Corrected r[b]
Operations	.45	.60
Combat Arms	.47	.62
Skilled Technical	.58	.67
Electronics	.27	.34
All	.42	.54

[a]Correlation between knowledge and hands-on test scores averaged across samples.

[b]Correlation between knowledge and hands-on test scores averaged across samples and corrected for attenuation.

As in previous research, the correlations observed here indicated that knowledge tests and performance tests are highly related, but should not be freely substituted. In general, correlations were at a level consistent with those found in the literature. They were particularly high for three MOS, 64C Motor Transport Operator, 11B Infantryman, and 71L Administrative Specialist, that represented three separate occupational groupings. They were particularly low in two skilled technical occupations, 95B Military Police and 91A Medical Specialist. This pattern in the skilled technical grouping does not correspond to findings reported in the literature (Rumsey, Osborn, & Ford, 1985). Since the Military Police and Medical Specialist occupations were also the MOS for which scores on a cognitive qualifying test, the Armed Forces Qualification Test (AFQT), were higher than in any of the other Project A occupations examined, there is some reason to believe that restriction in range may have been a factor contributing to the rather low correlations found there.

How reliable were the measures developed here? Such a question must be approached with considerable caution. Internal consistency indexes are those typically reported in the literature for both hands-on and job knowledge tests; accordingly, we have reported such indexes here. Yet we must recognize their limitations. Internal consistency reliability is most appropriate when the objective is to measure factorially pure traits and items are mutually independent. The factorial purity of proficiency tests will vary according to the content of the job. As Tenopyr (1985) has noted, "Any problems with internal consistency may reflect only the fact that job tasks are not homogeneous."

The strategy adopted in Project A of trying to maximize job coverage by measuring at least one task in every cluster was one that would tend to produce relatively heterogeneous tests and depressed estimates of internal consistency. Furthermore, item independence was violated in hands-on tests when an examinee's failure to perform the initial steps of a task made it impossible for that examinee to perform the final steps as well.

The weighted average of the split-half reliability estimates shown in Table 6 for the 30 knowledge tests is .80. This average does not substantially deviate from an average reliability of .83 reported in the literature for job knowledge tests (Rumsey, Osborn & Ford, 1985).

The average of the split-half reliability estimates shown in Table 5 for the 15 hands-on tasks was .52. Ultimately, a 30-task test will be generated for each MOS based on the 15 tasks for which both types of measures have been developed and the 15 tasks for which only job knowledge tests have been developed. Using the Spearman-Brown formula, it can be estimated that the reliability of a 30-task hands-on test would have been .68, relative to an average value of .71 reported in the literature (Rumsey, Osborn, & Ford, 1985). While the internal consistency estimates found here were clearly not high relative to those previously reported, that fact should not be alarming given that the overall test development strategy emphasized comprehensiveness more than content homogeneity.

The particularly low internal consistency estimates for the individual task tests, both written and hands-on, are to a large degree a function of the limited length of such tests. Few of these tests are by themselves stable indicators of a soldier's performance; it is only when test scores are summed across an MOS that a reasonable degree of stability might be expected.

The high correlations among rating scales, relative to their correlations with other methods, are neither surprising nor disappointing. Not only are the rating scales a visibly different method for measuring task performance, but they are deliberately addressing an affective component of performance, rather than the technical skill and knowledge aspects measured by the task tests. Interrater reliabilities were sufficiently high and scientific interest sufficiently whetted to warrant retention of the scales for the Concurrent Validation.

Nevertheless, findings reported by Borman, White, Gast and Pulakos (1985), using this same field test data set, reveal that, for some MOS, overall performance ratings were more closely related to hands-on and job

knowledge tests than to the task-based ratings examined here. This raises questions about whether raters really adequately understood and appropriately used the task-based scales.

The results from the Job History Questionnaire, while far from conclusive, provided sufficient indication that job experience may be an important factor to warrant further consideration of this variable. As a consequence, the Job History Questionnaire is being retained in the Concurrent Validation data collection. Those data, with much larger sample sizes, will be used to identify which, if any, task measures should be corrected for the contaminating effects of differential experience. Furthermore, the relationship between experience and performance may vary as a function of the aptitude being validated and the difficulty of the task. Therefore, care will be taken regarding the possibility of interaction effects as well as covariance effects.

The developmental activities described in this report resulted in the preparation of performance measures to be administered concurrently with predictor measures in a large-scale testing effort. As this effort is completed, a new set of task-based measures will be developed to measure performance of soldiers in their second tour. It is anticipated that many of the procedures used in developing first-tour measures will be appropriate for this new purpose as well, but it is also anticipated that some revisions will be needed to accommodate the expanded responsibilities associated with second-tour jobs. Work on developing these revised procedures is already under way.

REFERENCES

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Borman, W. C., White, L. A., Gast, I. F., & Pulakos, E. D. (1985). Performance ratings as criteria: What is being measured? Paper presented at the conference of the American Psychological Association, Los Angeles.

Campbell, J. P. (1983). Sample selection. In N. K. Eaton & M. H. Goer (Eds.), Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the annual report (ARI Research Note 83-37). (AD A137 117) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Davis, R. H., Davis, G. A., & Joyner, J. N. (1985). Development and field test of job-relevant knowledge tests for selected MOS (in preparation). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Eaton, N. K., & Goer, M. H. (Eds.). (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1983 fiscal year (ARI Research Report 1347). (AD A141 807) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Eaton, N. K., & Goer, M. H. (Eds.). (1983). Technical appendix to the annual report (ARI Research Note 83-37). (AD A137 117) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (Eds.). (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (ARI Technical Report 660). (AD A178 944) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report synopsis, 1984 fiscal year (ARI Research Report 1393). (AD A173 824) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Appendices to annual report, 1984 fiscal year (ARI Research Note 85-14). (In preparation) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). (AD A156 807) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Peterson, N. (Ed.). (1985). Development and field test of the trial battery for Project A (in preparation). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Pulakos, E. D., & Borman, W. C. (Eds.). (1987). Development and field test of the Army-wide ratings scales and the rater orientation and training program (ARI Technical Report 716). (In preparation) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rosse, R. L., Borman, W. C., Campbell, C. H., & Osborn, W. C. (1983). Grouping Army occupational specialties by judged similarity. Paper presented at the conference of the Military Testing Association, Gulf Shores, AL. In N. K. Eaton & M. H. Goer (Eds.), Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the annual report (ARI Research Note 83-37). (AD A137 117) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985). Comparing work sample and job knowledge measures. Paper presented at the conference of the American Psychological Association, Los Angeles.

Tenopyr, M. L. (1985). Building a composite measure of soldier performance: Discussant's presentation. Paper presented at the conference of the American Psychological Association, Los Angeles.

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1985). Development and field test of behaviorally anchored rating scales for nine MOS (in preparation). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

U.S. Army. (1983, October 3). Soldier's Manual of Common Tasks (FM 21-2).

Vineberg, R., & Taylor, E. N. (1972). Performance in four Army jobs by men at different aptitude (AFQT) levels: 4. Relationships between performance criteria (HumRRO Technical Report 72-23). Alexandria, VA: Human Resources Research Organization (HumRRO).